Human Molecular Genetics, 2007, Vol. 16, No. 23 **2783–2794**
doi:10.1093/hmg/ddm208
Advance Access published on July 31, 2007

Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases

Adam J. de Smith[1,†], Anya Tsalenko[2,†], Nick Sampas[2], Alicia Scheffer[2], N. Alice Yamada[2], Peter Tsang[2], Amir Ben-Dor[2], Zohar Yakhini[2], Richard J. Ellis[3], Laurakay Bruhn[2], Stephen Laderman[2], Philippe Froguel[1,4] and Alexandra I.F. Blakemore[1,*]

[1]Genomic Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK, [2]Agilent Laboratories, Santa Clara, CA, USA, [3]North West London Hospitals NHS Trust Regional Genetics Service, Northwick Park Hospital, Harrow, UK and [4]CNRS 8090-Institute of Biology, Pasteur Institute, Lille, France

Received June 14, 2007; Revised and Accepted July 23, 2007

The discovery of copy number variation in healthy individuals is far from complete, and owing to the resolution of detection systems used, the majority of loci reported so far are relatively large (~65% > 10 kb). Applying a two-stage high-resolution array comparative genomic hybridization approach to analyse 50 healthy Caucasian males from northern France, we discovered 2208 copy number variants (CNVs) detected by more than one consecutive probe. These clustered into 1469 CNV regions (CNVRs), of which 721 are thought to be novel. The majority of these are small (median size 4.4 kb) and most have common boundaries, with a coefficient of variation less than 0.1 for 83% of endpoints in those observed in multiple samples. Only 6% of the CNVRs analysed showed evidence of both copy number losses and gains at the same site. A further 6089 variants were detected by single probes: 48% of these were observed in more than one individual. In total, 2570 genes were seen to intersect variants: 1284 in novel loci. Genes involved in differentiation and development were significantly over-represented and approximately half of the genes identified feature in the Online Mendelian Inheritance in Man database. The biological importance of many genes affected, along with the well-conserved nature of the majority of the CNVs, suggests that they could have important implications for phenotype and, thus, be useful for association studies of complex diseases.

INTRODUCTION

Genomic copy number variation is much more common and involves a much greater proportion of the genome than previously realized (1–10). In addition to the intrinsic interest of elucidating the structure, evolution and current variability of the human genome, research into copy number variants (CNVs) is important as they may contribute to susceptibility to common diseases. It is essential to characterize and catalogue genomic copy number variation in healthy individuals as a foundation for assessing its putative implications for disease-associated phenotypes relevant to common complex disorders. Particular CNVs are already reported to be associated with susceptibility to HIV, glomerulonephritis and autism (11–14). CNVs are also of great interest to clinical cytogeneticists, who need to know what variation is 'normal' in the human genome in order to be able to determine which submicroscopic aberrations may be responsible for the rare, but very abnormal, phenotypes in young individuals, which are known as genomic disorders (15).

The CNVs reported so far have been documented in the TCAG database of genomic variants (16), which at the time

*To whom correspondence should be addressed. Tel: +44-2083832366; Email: a.blakemore@imperial.ac.uk
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

of writing contains over 3600 CNV loci. The overlap between CNVs reported by the various studies is not large (17), implying that information on copy number variability remains significantly incomplete (18). This is likely to be due to a number of factors. First, the size distribution of CNVs detected is dependent on the technology used. Secondly, given the probable differences in frequency of particular CNVs in different populations, the variety of different CNVs observable in a given study may be significantly limited by the number and ethnic origin of individuals examined. Thirdly, the fraction of common CNVs (minor allele frequency $>5\%$) may not be large, i.e. there may be many CNVs in the population, of which most tend to be rare. Finally, the number of false-positive and false-negative CNVs in different studies may be variable, depending both on the measurement technology and on the sample quality, including the sample source. For example, many of the CNVs reported thus far have been discovered in DNA derived from cell-culture lines (4,6,8,9), which are known to be susceptible to genomic changes during propagation (19).

In this report, we add a substantial number of new variations to the growing list of CNVs and begin to address some of the issues listed earlier. We studied CNVs in a population of 50 apparently healthy, middle-aged Caucasian males of northern French origin, using genomic DNA derived from peripheral blood in order to avoid artefacts because of cell culture. We used high-resolution array comparative genomic hybridization (aCGH) with 60mer oligonucleotide probes, at an average spacing of 500 bp spanning 2475 regions of the genome identified in this study as putative CNVRs in this sample set. We also examined and contributed to the validation of 2148 intervals previously reported as CNVRs by determining the proportion that is variant in the study population.

We found 2208 CNVs detected by two or more probes (hereafter called multi-probe CNVs), clustering into 1469 CNVRs: of these, 721 CNVRs do not overlap regions already represented in the TCAG database of genomic variants. These 721 novel regions contain 367 genes, 150 of which are represented in the Online Mendelian Inheritance in Man (OMIM) database. The majority of CNVs that we observed are relatively small ($66\% < 20$ kb). We also found that the breakpoints of many CNVs are highly conserved between individuals, supporting the possibility that they might have high utility for association studies.

RESULTS

We employed two different array designs for this study. The first was a 185K feature genome-wide scanning array, whereas the second was a focused custom 244K feature array designed to measure a large set of putative and known CNVs at high resolution. Throughout this article, single-probe intervals describe those identified by a single probe on the array, and multi-probe intervals are those identified by two or more probes. Although it is often customary to report only intervals detected by more than one consecutive probe, there is evidence from our results that a significant proportion of the single probe intervals called in this study represent real events, as outlined below. We have, therefore, included single-probe intervals in our analysis of the characteristics of the putative CNVs identified in this sample set.

In the first phase of the study, DNA derived from blood of a random subset (35 samples) of the 50 apparently healthy, middle-aged white males of northern French origin was hybridized (in two-colour experiments with a pool of the 50 samples as the reference sample) to genome-wide microarrays comprising 185 000 60mer oligonucleotide probes. These probes were designed for CGH, with an average spacing of 16 kb, but with probe placement biased towards genes. Using the aberration detection module (ADM)-1 algorithm (20) with a threshold of 6 (relaxed stringency to ensure effective capture of putative CNVs), 1003 multi-probe variant and 3777 single-probe variant intervals were detected (Fig. 1A, Supplementary Material, Fig. S1A and Materials and Methods). Consistent with data from previous studies, variations detected using the genome-wide scanning arrays were distributed throughout the genome with the number of CNVs detected roughly proportional to the size of the chromosomes (Supplementary Material, Fig. S2). We selected 2475 putative CNVRs for further investigation. These fell into three main categories as follows: all of the 1093 intervals that were either called in two or more samples by CGH Analytics common aberration analysis (21) with ADM-1 threshold of 6 or were found to have bimodal distributions of $\log_2$ ratios across the 35 samples; 729 putative CNV intervals each found in a single sample and a subset (653) of the regions that either exhibited trimodal distribution of $\log_2$ ratios across the 35 samples or were identified by CGH Analytics common aberration analysis with a less stringent ADM-1 threshold of 4.

For the second phase of the study, we designed a focused custom 244K feature array with a higher density of probes (spacing from 500 to 1500 bp) within and flanking 2475 putative CNVRs from the first phase (described earlier) as well as probes in 2148 intervals previously recorded as CNVRs in the October 2006 version of the TCAG database of genomic variants (average spacing of 5 kb). Genomic DNA, from the 50 samples, was hybridized in two-colour experiments using an individual DNA sample (NA15510) as the reference. Analysis of the phase 2 data, using the ADM-2 algorithm (20,22) with threshold 4, identified a total of 9244 multi-probe CNV intervals in the 50 samples, with a mean of 197 multi-probe intervals in each individual (Fig. 1B, Supplementary Material, Fig. S1B), compared with a mean of 10.3 in control self–self hybridizations (as described in Materials and Methods). The minimum number of multi-probe CNV intervals in any sample was 127 and the maximal number was 362 (Fig. 2A). In addition to the multi-probe CNVRs, there were also 6089 putative CNVs detected by single probes, 48% of which were observed in more than one individual (Supplementary Material, Table S1). There was a median of 669 single-probe intervals in each individual (compared with a mean of 29.6 in control self–self hybridizations) (Fig. 2B).

The 9244 multi-probe variant intervals found in the 50 individuals were grouped into CNVRs, as described in Materials and Methods and in Supplementary Material, Figure S3. We detected a total of 1469 multi-probe CNVRs, of which 1064 regions were definitely greater than 1 kb and 405 regions were detected by two or more probes less than 1 kb apart, although their exact size was not defined (Table 1 and

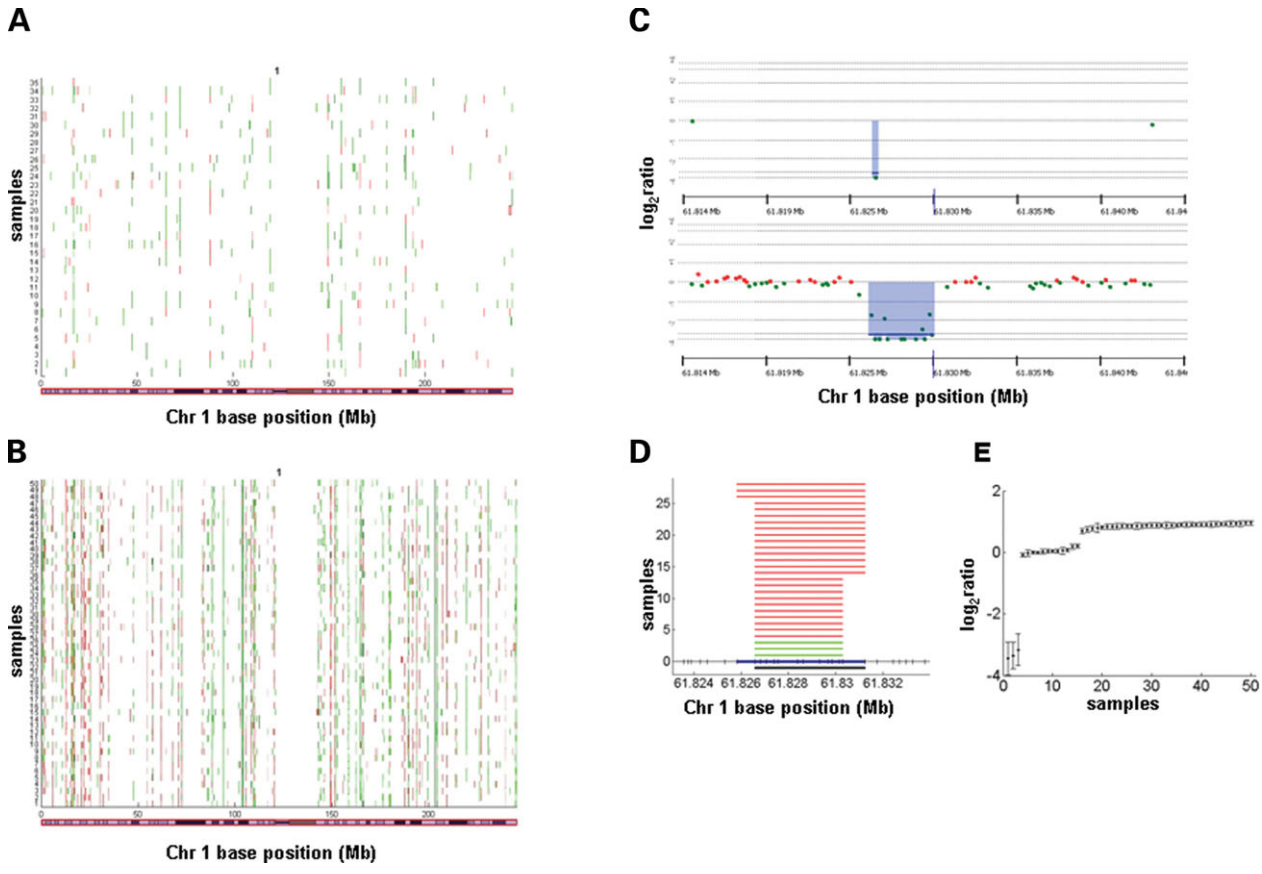

Figure 1. CNV intervals identified on chromosome 1. (**A**) Putative CNV intervals on chromosome 1 from the 185K genome-wide scanning array plotted as a function of chromosomal position for each of the 35 individuals. CNV intervals in each individual were identified by the ADM-1 algorithm with a threshold of 6. Each row shows CNV intervals in one person, red bars indicate gains when compared with the reference sample, whereas green bars indicate losses compared with the reference sample. Vertical alignment of gains and losses indicates positions where different individuals have copy number variation in the same region. (**B**) CNV intervals on chromosome 1 from the 244K custom-focused array are plotted as a function of chromosomal position for each of the 50 individuals. CNV intervals in each individual were identified by the ADM-2 algorithm with a threshold of 4. (**C**) Detailed view of CNVR on chromosome 1 near position 61.8 Mb as defined by genome-wide 185K array (top) and focused 244K array (bottom). A copy-number loss compared with the reference identified in three samples (one shown) by a single probe from the genome-wide 185K array is shown in the top plot along with two flanking probes. A loss in the same sample was detected by 12 probes in the focused 244K array (bottom plot). (**D**) Schematic representation of per-sample CNVs called in chr1: 61,825,811-61,831,251 bp CNVR. Red and green lines represent gains and losses in individual samples, respectively. Twenty-five samples have gains and three samples have losses. The corresponding CNVR is shown by a thick blue bar. CNVR boundaries extend from the leftmost to the rightmost breakpoint of per-sample CNV intervals. The CNV that represents the median position of breakpoints of the per-sample CNV intervals is shown by the thick black bar. Probe positions are indicated as small marks on the 'zero' line. Per-sample breakpoints of CNV intervals vary by only one probe. (**E**) Mean and error on the mean of $\log_2$ ratios of probes in the chr1: 61,825,811-61,831,251 bp region from the 244K custom-focused array across all 50 individuals. Means of $\log_2$ ratios cluster in three different groups corresponding to different copy-number levels. Levels of copy-number loss in three samples (mean $\log_2$ ratio < -3) indicate that these are homozygous deletions when compared with the reference. The number of samples with gains, relative to the reference, and the absence of samples with $\log_2$ ratios consistent with hemizygous loss indicate that the reference sample itself may be hemizygous in the region, and the 25 samples showing gains have two copies.

Supplementary Material, Table S2). A breakdown of the frequency distributions for these CNVRs is shown in Supplementary Material, Figure S4. The median size of CNVRs was 4.4 kb, and large variations (of size more than 100 kb) account for only 12% of CNVRs identified in this study, compared with 45% of variations reported in the TCAG database (Fig. 3). Of the 6089 single-probe intervals discovered, a proportion is contained inside larger variations observed in other people, whereas 4705 are all outside of the multi-probe CNVRs identified in this study.

Novel regions

To assess what proportion of the variant regions discovered in this study was novel, we compared our data with those in the TCAG database (March 2007) (16). In this study, 748 multi-probe CNVRs (51%) overlapped loci reported in the TCAG database: 518 of these regions were completely contained within the corresponding TCAG variants and, therefore, have size smaller than previously reported based on this cohort. Another 721 regions are not represented in the TCAG database. Of these 721 novel regions, 416 have sizes larger than 1 kb, and only 34 regions are larger than 100 kb (Fig. 3). Of the 4705 single-probe intervals outside the multi-probe CNVRs identified, 2662 are potentially novel and do not overlap with regions reported in the TCAG database (Table 1).

Similarly, we investigated what proportion of the total number of previously reported CNVRs was detected in this population. The multi-probe CNVRs identified in our study

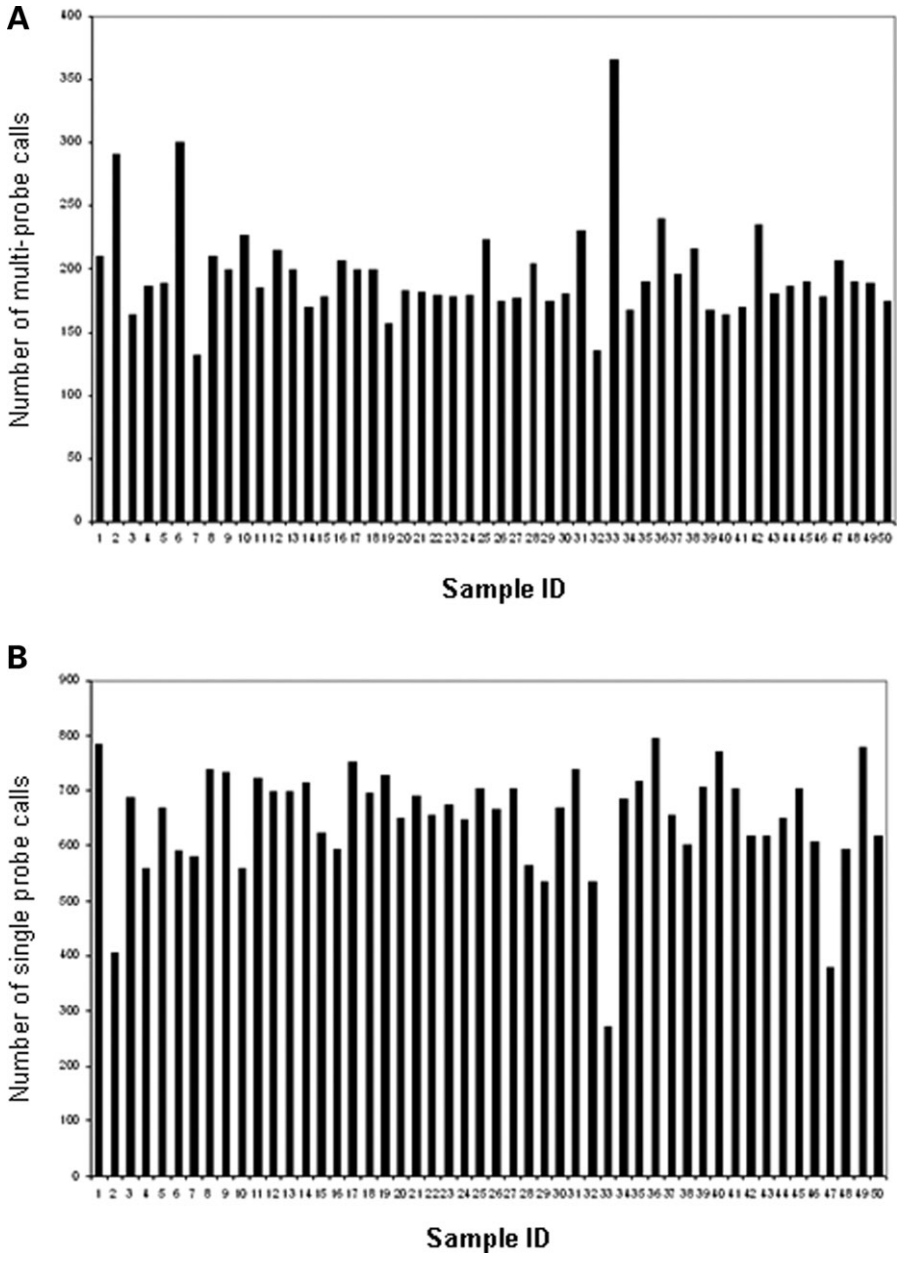

Figure 2. Number of putative CNV intervals called in each of the 50 French samples. The number of multi-probe (**A**) and single-probe (**B**) CNV intervals called using the ADM-2 algorithm with threshold 4 is plotted for each of the 50 samples.

overlapped 1735 of 6560 (26%) variations reported in the TCAG database (March 2007) (16).

Confirmation and refinement of CNVs

Over 95% of the putative multi-probe CNVRs seen in two or more individuals using the phase 1 genome-wide 185K array were subsequently called as CNVRs using the phase 2 focused confirmation array. Of the common and bimodal single-probe CNV intervals from phase 1 (ADM-1, threshold 6), 60% were validated as multi-probe calls by the focused phase 2 array data and an additional 30% of the phase 1 calls were again called by a single probe in the second phase. An example of a single-probe call from phase 1 data confirmed in phase 2 by 12 probes is shown in Figure 1C.

We have also provided independent confirmation of 551 loci and 1735 variations in the TCAG database. In addition, we have provided detailed information on the boundaries of CNVs in the studied population (Supplementary Material, Table S3), which contributes to the refinement of mapping positions and boundaries of many variants. For example, in analysis of a CNVR that encompasses the *FCGR3B* gene, which has been shown to be associated with predisposition to glomerulonephritis (12,23), we were able to provide additional information about the extent and the structure of

Table 1. Summary of CNVRs identified by focused 244K array in 50 samples

	Regions identified by multiple probes		Regions identified by a single probe	
	All CNVRs	CNVRs observed in multiple samples	All CNVRs	CNVRs observed in multiple samples
Total number	1469	660	4705	2057
Novel (do not overlap with TCAG variations)	721	269	2662	1168
Contain genes	726	349	1714	751
Novel and contain genes	350	131	984	444

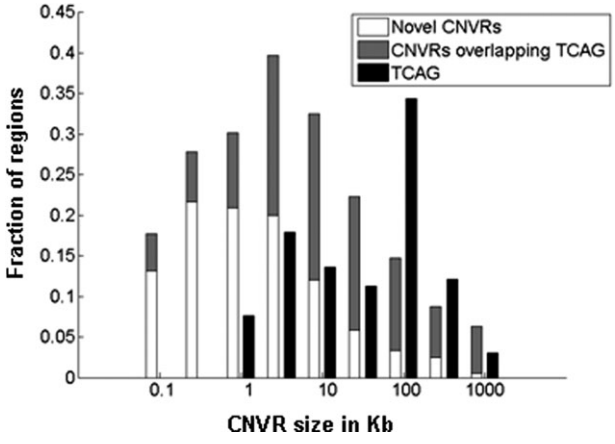

Figure 3. Comparison of distributions of sizes of CNVRs found in this study and of sizes of variations in the TCAG database as of March 2007. The fraction of CNVRs found in this study (white and grey bars) and variations reported in the TCAG database of genomic variations (black bars) are plotted for each of the size ranges indicated on the *x*-axis. Each bar represents the fraction of regions within size ranges centred at 10-fold multiples of 1 and $\sqrt{10}$, from 0.1 to 1000 kb. The CNVRs found in this study are divided into two groups: novel CNVRs (white) and CNVRs overlapping variations reported in the TCAG database (grey). The median size of CNVRs found in this study is 4.4 kb, and 405 regions are smaller than 1 kb. The March 2007 version of TCAG database used for this analysis does not contain variations smaller than 1 kb.

this CNV more definitively than in previous studies. We identified three CNVs (see Materials and Methods) in this region. Two variants showed only losses compared with the reference, a smaller variant of size 83.4 kb was observed in 18 samples and a larger variant of size 511.8 kb was observed in 16 samples (Fig. 4C). The third, very small, variant (detected by two probes 238 bp apart) was only observed in one individual and showed gain when compared with the reference. We confirmed that, in addition to *FCGR3B*, the most common CNVR encompasses five other genes, *FCGR3A, FCGR2A, FCGR2B, FCGR2B* and *HSPA6*, which might contribute to the observed phenotypic associations with autoimmune disease (Fig. 4A). The larger variant also contained *FCRLA, FCRLB, DUSP12, FCRLM1, FCRLM2, ATF6, OLFML2B* and *NOS1AP* genes. We have previously published microarray data showing good agreement between $\log_2$ ratios from individual probes in this region with copy-number estimations derived from quantitative polymerase chain reaction (PCR) results at *FCGR3B*, but confident assignment of individuals to discrete copy number classes on the basis of PCR has been problematic (23). Using mean $\log_2$ ratios from the 168 probes across this region, it is possible to identify four distinct copy number states: multiple copy loss, single copy loss, copy number equivalency and gain relative to the reference sample (Fig. 4B). All samples with the larger variation were in the single copy loss cluster, whereas those with the smaller variation were in both the single and multiple copy loss clusters.

In addition to confirmation of previously reported variants by aCGH, we have validated 21 variant loci from this study using alternative methods, such as PCR across deletion breakpoints and detection of duplication by multiplex probe ligation amplification (MLPA) (Fig. 5). For example, we have confirmed a multi-probe deletion by PCR in 14 samples within the *AKAP13* gene (a kinase anchor protein 13 = lymphoid blast crisis oncogene), showing both homozygous and heterozygous states in different samples (Supplementary Material, Fig. S5). In addition, we observed a novel 1.1 Mb duplication on chromosome 7q34, encompassing four genes (*PTN, DGK1, CREB3L2* and *AKR1D1*). Analysis of this region using MLPA on two daughters of the subject and their mother revealed that the two daughters inherited the copy number gain from the father (Fig. 5). Despite having inherited an extra copy of all four genes, the two daughters were also apparently healthy.

Properties of multi-probe CNVRs

We next examined the degree to which different variants within the CNVRs observed in this study vary with regard to the length and position of the CNVs and/or their copy number states between different individuals. All the individual CNV intervals observed were divided into a total 2208 CNVs (Supplementary Material, Table S3 and Fig. S3), assigning intervals observed in different individuals to the same CNV if they overlap by more than 50%, as described in Materials and Methods. Some regions of the genome show a high degree of complexity, with clusters of overlapping variants (Supplementary Material, Fig. S6) and others appear less complex, having only one variant (for example, the region shown in Fig. 1D). Of 660 CNVRs observed in more than one sample, 387 regions had only one variant and 115 CNVRs had two or more variants, observed in two or more samples each.

The great majority of variants had common boundaries: the coefficient of variation of endpoints in the multi-probe CNVs observed in multiple samples is less than 0.1 for 83% of the

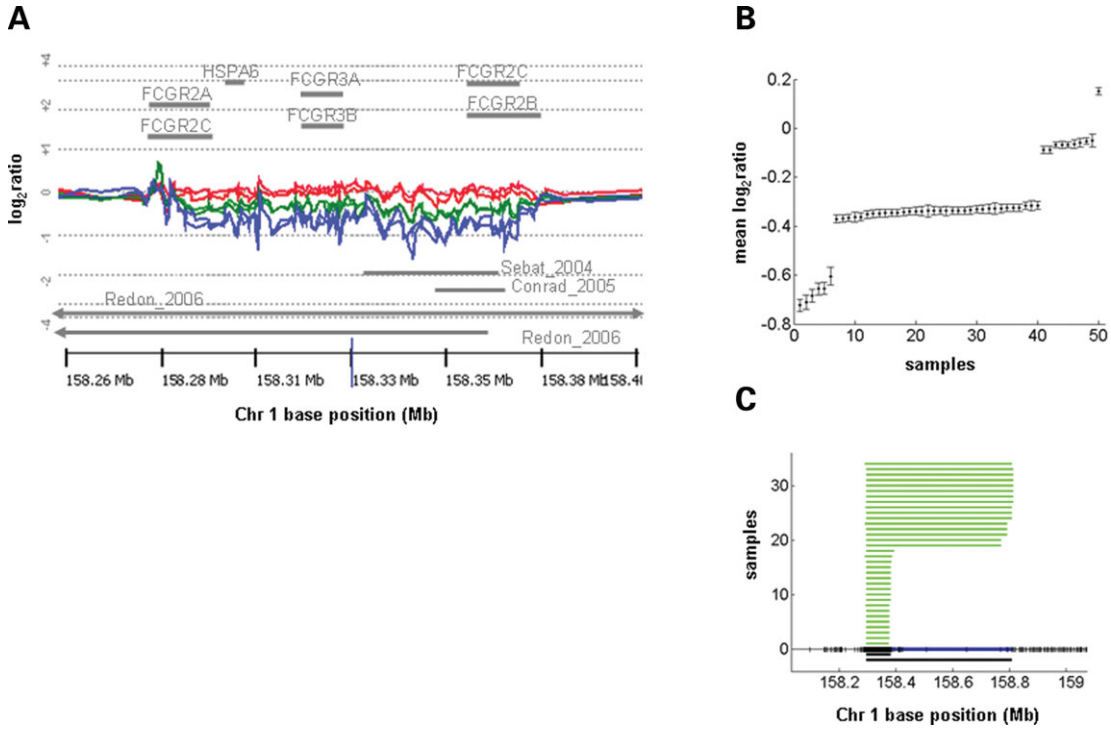

Figure 4. Detailed characterization of the CNVR on chr1: 158,291,736–158,811,985 bp (1q23.2), which encompasses the *FCGR3B* gene. (**A**) Log_2 ratios for six samples are plotted using a three-point moving average as a function of their chromosomal position. The CNV detected in these samples includes the *FCGR3B* gene, as well as *FCGR3A*, *FCGR2B*, *FCGR2C* and *HSPA6* and overlaps *FCGR2A*. Gene positions are annotated above, previously described copy number variations below according to the various publications. The six samples shown here are colour-coded by mean ratio across the region of loss. (**B**) Mean and error on the mean of log_2 ratios of the 165 probes within the region are plotted for each of the 50 samples. Mean values of log_2 ratios cluster in four different groups corresponding to different copy-number states.(**C**) Schematic representation of per-sample CNVs called in chr1: 158,291,736–158,811,985 bp CNVR. Green lines represent losses in individual samples (detected gain is not shown). CNVR is shown by a thick blue bar. CNVR boundaries extend from the leftmost to the rightmost breakpoint of per-sample CNV intervals. Two CNVs (shown by thick black bars) with observed losses were identified in this region: a smaller variant of size 83.4 kb was detected in 18 samples and a larger variant of size 511.8 kb was detected in 16 samples. The third variant of size 238 bp showing gain when compared with the reference is not shown. Probe positions are indicated as small marks on the zero line.

endpoints (Fig. 1D). Variants tended to overlap by more than 90% if they overlapped at all. The complete distribution of overlap of multi-probe CNVs is shown in Supplementary Material, Figure S7. Moreover, of the 1776 breakpoints of CNVs defined by multiple samples, 1142 endpoints were exactly the same, i.e. they ended at the same probe, in all samples that had copy number gain or loss at that position.

In the majority of multi-probe CNVRs, we observed only gains or only losses when compared with the reference sample (rather than both gains and losses observed at the same locus). Specifically, 39% of variants showed only gains (compared with the reference sample), whereas 55% showed only losses. A mere 6% of variants had both gains and losses. Forty-two variants were determined to be homozygous deletions in the reference sample based on fluorescent signal distributions (see Materials and Methods).

To determine the proportion of multi-probe CNVs showing multiple copy number states (e.g. putative hemizygous and homozygous deletions, or the presence of multiple levels of copy-number gain), we further analysed the relative frequency of different copy number states in 888 CNVs that were observed in three or more samples. In the majority of these CNVs, only one alternative copy-number state was observed. Of the regions with observed losses, 27 had two distinct peaks in the distribution of average log_2 ratios in samples with losses ($P < 10^{-6}$), indicating two distinct copy number loss events when compared with the reference sample (for example, the FCGR region shown in Fig. 4B) and 294 CNVs had only one peak in average log_2 ratios corresponding to losses (for example, the chromosome 1 region shown in Fig. 1E). Eight regions had two distinct peaks in the distribution of average log_2 ratios in samples showing gains ($P < 10^{-8}$), indicating two distinct copy-number gain events when compared with the reference sample, compared with 235 regions having only one peak in the distribution of average log_2 ratios in the samples with gains (for example, the chromosome 1 region shown in Fig. 1E).

Genes in CNVRs

In CNVRs identified by multiple probes, we found 1653 unique genes: 368 (22%) of which were in novel CNVRs, whereas 1286 (78%) were in CNVRs already described in the TCAG database. In addition, we found a total of 1302 genes affected by the 6259 single-probe putative CNVs in our population. Of these, 386 genes were also affected by multi-probe CNVRs (Supplementary Material, Fig. S8). For 282 genes, exonic sequences were included: 129 of these

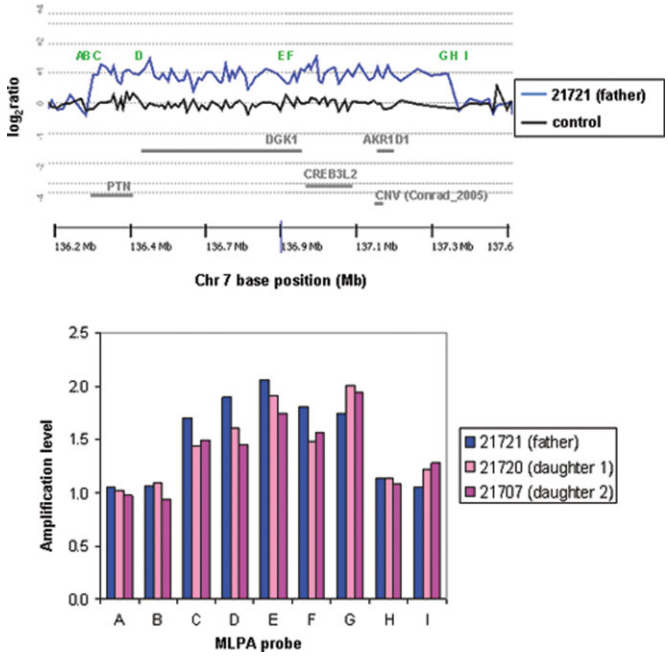

Figure 5. MLPA validation of ~1 Mb gain on chr7: 136,377,268–137,425,104 bp (7q34) in sample 21721 and his two daughters. (Top) MLPA target probes were designed to genomic sequences: just upstream (A) or downstream (I) of the probes flanking the amplified region; just downstream (C) or upstream (G) of the terminal probes included in the region and within the region of gain (D–F). Probes B and H were designed to sequences between the CGH array probes that define the boundaries of the amplified region. Ratios of the normalized MLPA amplification level relative to the unamplified mother are plotted in the bar chart (bottom). The four probes falling outside the region of gain (A, B, H and I) show a ratio of approximately 1, whereas those in the amplified region (C–G) confirm the gain in all three individuals.

genes showed variation at the same location in at least two people in our population. For genes with single-probe variations outside the exons, 635 of 1020 genes were affected in at least two people.

In order to investigate the potential biological consequences of the copy-number variation that we observed in this study, we conducted a gene ontology (GO) analysis, which identifies over- and under-represented categories for the genes in CNVRs. We generated three gene lists for comparison, consisting of genes in (i) TCAG CNVRs detected in this study, (ii) novel CNVRs and (iii) all CNVRs observed in the study. In all three cases, association with the plasma membrane was identified as a highly significant category (Supplementary Material, Table S4). We eliminated 'cellular component' from the analysis to reduce the redundancy and to highlight the biological function for genes in the various categories of CNVRs (Table 2).

Although the same GO categories were frequently represented in all three GO result sets, differences were observed (Table 2 and Supplementary Material, Table S4). For example, the GO categories with the highest significance for over-representation in the TCAG gene list are associated with sensory perception, whereas more general categories of differentiation and development top the GO categories of the genes in novel CNVs. In particular, nervous system development showed strong over-representation, along with functionally similar GO categories, such as transmission of nerve impulse and brain development (which comprised a different set of genes than those in nervous system development) (Table 2). The majority of the differences between the TCAG GO results and the GO results from all CNVRs observed was driven by single-probe CNVs: in genes covered by the TCAG intervals, elimination of the 215 genes affected by single-probe CNVs had minimal effect on GO representation. Similarly, elimination of the 1302 genes with single-probe CNVs from the list of genes in all observed CNVRs generated GO results that were very similar to the TCAG GO results (Table 2 and Supplementary Material, Table S4). Statistical significance of the GO over-representation associated with our novel CNVs was also highly influenced by the single-probe CNVs; the subtraction of the 197 genes in multi-probe CNVRs had minimal effect on the GO results in novel CNVRs (Supplementary Material, Table S4). As might be expected, the GO results from all CNVRs reflect contributions from GO categories associated with both TCAG and novel CNVRs.

Almost half (150 of 368) of the genes in novel multi-probe CNVRs were represented in the OMIM (24), suggesting that many genes in variant regions are likely to have disease relevance and/or biological importance. We observed instances where individuals in our sample population appeared to lack one of the copies of biologically important genes, such as the lipoprotein lipase gene (*LPL*), involved in receptor-mediated lipoprotein uptake (25), and the carbonyl reductase 3 gene (*CBR3*), important in metabolizing pharmacologically relevant carbonyl compounds (26). We also found instances where individuals contain copy-number gains of regulatory genes, such as *NKX2-8*, which encodes an nk2-related transcription factor associated with hepatocellular carcinomas (27). Genes encoding another family of transcription factors known as the KRAB-Kruppel zinc fingers were also found in variant regions (including *ZNF12, ZNF14, ZNF441, ZNF536* genes as well as closely associated *ZNF165* and *ZNF696*); these transcription factors are strong repressors that are evolutionarily unstable, with a high degree of copy number variability between species studied (28). Approximately half (634) of the genes that overlap single-probe variants anywhere in the gene were also represented in OMIM (of which 73 had exonic variants that were found in more than one person). Within this list is an assortment of genes of interest to human health, such as the cancer-relevant *MEN1, APC* and *IGFIR* as well as genes potentially associated with other complex phenotypes, including congenital malformations, neurological and psychiatric disorders.

DISCUSSION

We report discovery of a large number of novel copy number variations in the genomes of healthy adults. Many of these affect genes and, thus, may have important implications for subtle, quantitative or late-onset phenotypes. The measurement platform that we have used is designed to enable detection of small as well as large CNVs, allowing us to

Table 2. Over-represented GO categories associated with CNVRs in the sample population

	Group	Total	*P*-value	GO category
TCAG CNVRs (1286 genes)	57	525	6.7×10^{-30}	Sensory perception of smell
	57	619	6.5×10^{-23}	Olfactory receptor activity
	90	1620	1.5×10^{-12}	G-protein-coupled receptor protein signalling pathway
	12	24	1.8×10^{-11}	Pancreatic ribonuclease activity
	86	1623	1.9×10^{-10}	System development
	108	2289	2.2×10^{-9}	Multi-cellular organismal development
	15	64	5.1×10^{-9}	Epidermis morphogenesis
	12	39	1.2×10^{-8}	Keratinization
All CNVRs (2570 genes)	275	2289	1.2×10^{-38}	Multi-cellular organismal development
	102	703	1.0×10^{-21}	Nervous system development
	41	228	3.3×10^{-13}	Central nervous system development
	112	1008	8.7×10^{-12}	Cell adhesion
	143	1444	2.6×10^{-10}	Calcium ion binding
	207	2386	8.5×10^{-9}	Cell surface receptor linked signal transduction
	45	325	3.4×10^{-8}	Transmission of nerve impulse
	52	400	3.7×10^{-8}	Magnesium ion binding
Novel CNVRs (1284 genes)	242	3870	2.0×10^{-29}	Multi-cellular organismal process
	167	2289	2.0×10^{-29}	Multi-cellular organismal development
	157	2389	6.3×10^{-21}	Cell differentiation
	66	703	4.6×10^{-19}	Nervous system development
	32	228	3.0×10^{-18}	Central nervous system development
	79	1019	2.5×10^{-15}	Anatomical structure morphogenesis
	34	325	7.4×10^{-12}	Transmission of nerve impulse
	23	187	1.8×10^{-10}	Cell projection organization and biogenesis
	23	187	1.8×10^{-10}	Cell part morphogenesis
	50	642	1.2×10^{-9}	Transcription from RNA polymerase II promoter
	38	429	1.2×10^{-9}	Regulation of transcription from RNA polymerase II promoter
	89	1444	1.4×10^{-9}	Calcium ion binding
	29	286	1.6×10^{-9}	Synaptic transmission
	19	162	6.9×10^{-8}	Ubiquitin–protein ligase activity

Parent categories are shown in Supplementary Material, Table S4.

systematically study the size distributions of a large number of previously identified and novel CNVs in more detail than many previous studies. Twenty-eight percent of the CNVRs detected in our sample set were detected by probes less than 1 kb apart, suggesting that a larger proportion of copy number variation between individuals may reside in smaller intervals than previously reported. Outside these CNVRs, an additional 4705 variations were identified by single probes, 1587 of them even when probe spacing was as close as 500 bp. Although we have excluded probes that may contain restriction cut sites created by known single nucleotide polymorphisms (SNPs) (as explained in Materials and Methods) from our analysis, it is possible that some of the single probe calls could be related to sequence variants, rather than variation in copy number. We did, however, observe a high confirmation rate of regions that had been identified by single probes in the first phase of the study in the second stage: 60% of single-probe intervals were converted to multi-probe intervals and 30% were detected again by a single probe. It is clear that the high resolution of the second stage array (one probe/500 bp in some regions) has allowed the detection of a large number of smaller CNVs. What is not yet clear is how many of the putative copy number variations that remained as only called by single probes in stage 2 of our study reflect the presence of even smaller CNVs/indels: an unknown proportion may result from other types of local sequence variation.

Although a substantial number of novel CNVs have been revealed in this study, a relatively small number of subjects were included and they originated from one geographical region. There is also a relatively modest overlap between regions found in this sample set and regions found in previous studies. For example, the regions we identified overlap with 30% of variations reported by Redon *et al*. (6), 35% of variations reported by Conrad *et al*. (1) and 35% reported by Tuzun *et al*. (9). As such, our work represents an early stage in high-resolution detection of CNVs—similar studies should be carried out in larger sample sets incorporating a wider range of populations, particularly those of African origin, which have shown significantly greater genetic diversity than other groups (29,30).

This is one of the highest resolution genome-wide studies to date, with the custom array having probes located only 500 bp apart in many regions. This has allowed us to more precisely refine the mapping of the boundaries of many known and novel CNVRs. A large proportion of the CNVs we measured appear to have common boundaries between individuals. There is, however, a variety of more complex cases. In these instances, it appears that intervals of different sizes mapping to the same region of the reference genome are variant

between individuals. Specific study of complex regions will be necessary to elucidate the genome organization and structure of the sequences within these regions.

In addition to refining our understanding of structural diversity between individuals, more precise knowledge of the breakpoints of CNVs enables a more accurate accounting of the genes and regulatory regions impacted by copy number variation. Many genes are affected by copy number variation, and many such variants are at high frequency in the studied population. Although the exact functional consequences of the CNVs overlapping genes remain to be clarified, it is of great biological interest that such a large number of genes may differ in individual copy number counts for the full gene or have exonic sequences missing or amplified. It is entirely unknown whether these genes form functional products with missing or amplified regions or whether the polymorphic variants result in gene inactivation. In addition to this exonic variation, some variants in intronic and regulatory regions are likely to influence gene transcriptional activity.

The potential effect of CNVs on human phenotype has important implications for genomic evolution (31,32). Analysis of areas of the human genome that have been subject to recent selection has highlighted several GO categories that are strongly over-represented, including chemosensory perception and olfaction, acquired and innate immunity, gametogenesis, spermatogenesis, fertilization, metabolism of carbohydrates, lipids and phosphates and vitamin transport (33). Nguyen *et al*. (34) recently demonstrated that a subset of human CNVs have been retained in the population because of positive selection. Chemosensation and immune response genes have a well-documented role in adaptation to novel environmental niches (34), and reports have shown that genes involved in fertility and reproduction are subject to rapid adaptive evolution in primates because of sexual competition and defence against pathogens (33): all these gene groups are over-represented in CNVs. Additionally, we found strong over-representation of genes involved in development (particularly of the nervous system) and differentiation. Evidence from other sources supports the importance of large-scale chromosomal alterations in human evolution: for example, a common inversion at chromosome 17q21.31 is thought to be under positive selection in Europeans, having been associated with recombination rate and with increased numbers of children (35). In the case of CNVs, selective pressure can be observed in current human populations, for example, *CCL3L1* variation is associated with susceptibility to HIV/AIDS and *CYP2D6* variations affect drug metabolism (resulting in variation in effective dose and rate of adverse effects), respectively (11,36).

The same plasticity of the human genome that has contributed to its evolution may also result in the formation of detrimental genomic mutations (37). There is evidence for this with respect to the 17q21.31 inversion, mentioned earlier, where microdeletions in the same region lead to a mental retardation syndrome (38). The results of our study will be of interest to clinical cytogeneticists hoping to use aCGH technology to uncover cryptic chromosomal imbalances in patients. Conventionally, smaller aberrations and, particularly, single-probe signals have been regarded as artefactual — we provide evidence that in many cases, these may be real, common and affect important genes. Similarly, large variants affecting multiple genes have been regarded as more likely to be pathogenic. Although this may generally be true, the fact that a >1 Mb duplication encompassing four genes is present in the genome of an apparently healthy individual and his two daughters highlights the necessity of documenting all CNVs discovered. Cytogeneticists need to know which are likely to be phenotypically neutral in order to assess which aberrations may be responsible for congenital malformations and learning disability in patients with suspected genomic disorders.

Genomic disorders are rare, but CNVs, particularly those that are common and affect coding sequences or regulatory regions, may have wider implications for human health, with subtle variations in phenotype having important consequences for complex disease, such as cancer, neuropsychiatric diseases, obesity and diabetes. This concept is exemplified by the association of copy number polymorphism in FCGR genes with organ-specific immunity (12,23) and associations of other CNVs with autism (13,14) and with host response to HIV (11). Our study provides a significant refinement to the existing map of genetic differences and potentially affected genes. A more complete map of CNVs is a critical step in understanding the biological relevance of human genomic variation and evaluation of the contribution of CNVs to common diseases.

MATERIALS AND METHODS

DNA samples

DNA samples were isolated from peripheral blood of 50 unrelated, apparently healthy white males of northern French origin using Puregene kits (Gentra, USA) and resuspended in Tris–EDTA buffer. For analysis of Mendelian transmission, DNAs from three family members (spouse and two daughters) of one particular subject were also purified from peripheral blood — these subjects were also apparently healthy Caucasians from northern France. For the first phase of the study using the genome-wide 185 K array, a pooled reference sample was generated by combining an equal mass of genomic DNA from all 50 subjects. After phase 1 of the study, it became clear that many more CNVs than anticipated were common in the population and, thus, there would be loss of power to detect them using a pooled reference. For this reason, it was considered preferable to switch to a single reference for the second phase. The particular sample chosen, obtained from the Coriell Cell Repository, was derived from a north American female of unknown ethnic origin (NA15510). This sample has been extensively characterized and is recommended for use in CNV detection programmes to allow meaningful comparison of data between studies (discussed in Scherer *et al*. (39). All samples had Ethics Committee approval for use in this study.

Microarray design

Microarrays used in this study were 60mer *in situ* synthesized oligonucleotide arrays designed and produced by Agilent Technologies (Santa Clara, CA, USA). Two array designs

were employed. The first was a genome-wide CGH array consisting of 185 000 probes nominally spaced evenly across the genome with average probe spacing of 16 kb, but with a bias towards known genes. The second array was a custom focused array containing 244 000 probes selected from Agilent's High-Density database of over 8 million validated CGH probes. Probes in this database cover exonic, intronic and intergenic regions of the genome and have unique representation in the NCBI35 build of the human genome sequence. The custom 244K array has increased probe density within and flanking 2475 putative CNVRs from the first phase of this study as outlined below (at a spacing of 500–1500 bp) and in 2148 intervals reported in the October 2006 build of the TCAG database of genomic variants (at 5 kb probe spacing). This represents all but 43 of the CNVs included in the database at that time: these 43 regions were not covered by the candidate probe set used for this study. The putative CNVRs from the first phase covered on the focused confirmation array included: all the 235 multi-probe and 335 single-probe intervals that were called in two or more samples by CGH Analytics context-corrected common aberration analysis (21) with ADM-1 threshold of 6, and 523 additional single-probe intervals that were found to have bimodal distributions of $\log_2$ ratios across the 35 samples.

We had lower initial confidence in calls detected at lower thresholds of significance. To investigate these calls, we included a random sampling of 453 intervals identified by CGH Analytics context-corrected common aberration analysis with a reduced ADM-1 threshold of 4 and also sampled 200 single-probe intervals that exhibited trimodal $\log_2$ ratios distributions across the 35 samples. In addition, we included 729 calls, each observed in only one of 19 individuals. For these 19 individuals, all calls were included in phase 2. These probes were added to explore different categories of regions detected on the genome-wide array, but not necessarily observed in a significant number of samples or with a more stringent threshold.

Microarray labelling and hybridization

All array hybridizations were performed according to the manufacturer's recommended protocols (40). Briefly, 500 ng of genomic DNA was digested with restriction enzymes *Alu*I and *Rsa*I and fluorescently labelled using the Agilent DNA Labelling kit. *t*-test samples were labelled with cyanine 5-dUTP and the reference sample with cyanine 3-dUTP. Labelled DNA was denatured and pre-annealed with Cot-1 DNA and Agilent blocking reagent prior to hybridization for 40 h at 20 r.p.m. in a 65°C Agilent hybridization oven. Standard wash procedures were followed. Arrays were scanned at 5 μm resolution using an Agilent scanner, and image analysis was performed using default CGH settings of Feature Extraction Software 9.1.1.1 (Agilent Technologies).

Statistical analysis

Putative CNV intervals in each sample were identified using Agilent CGH Analytics 3.4 software (41) and Matlab-based tools using the same statistical methods and algorithmic approaches. All regions of statistically significant copy number change were determined using ADM algorithms (20,22). The ADM algorithms identify genomic regions with copy-number differences between the sample and the reference based on $\log_2$ ratios of fluorescent signals from probes in the interval. In brief, ADM algorithms use an iterative procedure to identify all genomic regions with the deviation of average of the measured signals in a given region from its expected value of 0 larger than a given threshold. This deviation is measured by a statistical score. At each iteration, the region with the most significant score is reported.

Throughout this article, single-probe intervals describe putative CNVs identified by one probe on the array, and multi-probe intervals are those identified by two or more probes. Boundaries and sizes of intervals are defined on the basis of positions of the last and the first microarray probes in the interval.

In order to capture a wide range of putative CNVs in the discovery stage, the genome-wide scanning array data were analysed using the less stringent ADM-1 algorithm at threshold 6. In a pre-processing step, features with log ratio error > 0.5 (in $\log_2$ scale) were filtered out. Centralization and fuzzy zero corrections were applied to remove putative variant intervals with small average $\log_2$ ratios. Aberrations common to two or more samples were identified by context-corrected common aberration analysis in CGH Analytics 3.4 (21). Bimodal probes were defined by the following criteria: probes were called bimodal if $\log_2$ ratios for this probe across 35 samples could be classified into two groups with significantly different averages as determined by a two-sample *t*-test ($P < 10^{-14}$). Trimodal probes were defined as probes that were not bimodal, for which the $\log_2$ ratios across 35 samples could be divided into three groups with significantly different averages by analysis of variance.

For the second phase, data from the focused confirmation array were analysed using the ADM-2 algorithm at threshold 4. The ADM-2 algorithm uses $\log_2$ ratios weighted by $\log_2$ ratio error as calculated by Feature Extraction software to identify genomic intervals with copy number differences between the sample and the reference. Data were centralized, and calls with average $\log_2$ ratios less than 0.3 were excluded from the analysis, as were any calls detected by probes containing a known SNP that may alter an *Alu*I or *Rsa*I restriction site as determined by the 9.3 million in the UCSC annotation database for the genome browser (42).

The false-positive rate for the ADM-2 algorithm at threshold 4 was determined using three self–self hybridizations of the reference sample. In the three replicate self–self experiments, ADM-2 analysis with threshold 4 identified an average of 29.6 single-probe intervals and 10.3 multi-probe intervals. Comparing the average number of variant interval calls in self–self experiments with the average number of variant interval calls for each sample, we estimated the false-positive rate to be 0.05 (10.3/197) for multi-probe calls and 0.04 (29.6/669) for single-probe calls.

The false-negative rate was estimated in a manner similar to that described by Wong *et al*. (10) based on four replicate experiments for one of the samples. In four replicate experiments, 223 putative variant intervals were observed two or more times and were considered true calls (49 intervals were

observed twice, 43 intervals were observed three times and 131 intervals were observed four times), yielding an estimate of false-negative rate of 0.16 [(2*49 + 43)/(4*223) = 0.16]. In this analysis, we conservatively considered aberrant intervals in two experiments the same if they overlapped by more than 0.9.

After putative variant intervals were identified in each sample, we used the following iterative procedure to determine the boundaries of CNVRs as illustrated in Supplementary Material, Figure S3. At each iteration, we identified a genomic locus *g* that was overlapped by the largest number of per-sample variant intervals. Each interval had to extend at least 250 bp to each side of the selected locus *g*. CNVR *R* corresponding to *g* was defined as the union of per sample intervals overlapping *g*. Per-sample intervals overlapping *g* and other intervals completely contained in *R* were excluded from the next iteration of the algorithm. Iteration steps were repeated until all per-sample intervals were combined into CNVRs.

Per-sample intervals corresponding to each CNVR were further clustered to define CNVs corresponding to each region. Each cluster contained per-sample intervals with pairwise overlap of at least 0.5. Boundaries of each CNV cluster were defined as median endpoints of its members.

Average $\log_2$ ratios of intervals showing gains and losses corresponding to each CNV were further analysed to identify distinct copy-number states. Using a two-sample Student's *t*-test, we identified intervals for which the average $\log_2$ ratios for gains/losses can be divided into two groups with significantly different means. In addition, a one-sample Student's *t*-test was used to test that the distribution of all average $\log_2$-ratios for gains/losses is significantly different from 0.

CNV intervals called in different samples can differ in the genomic intervals that they span as well as in the level at which their copy number differs from that of the reference sample. We analysed the data to determine the observed instances of both types of variations. We use the term CNV to refer to variants that differ in length, whereas CNVs that differ in copy number between individuals are referred to as having different copy-number states.

Homozygous deletions in the reference sample were determined on the basis of the absolute and relative signal intensities in the reference channel when compared with signals across all samples. Variants were determined to be homozygous deletions in the reference sample if either of the following was true. The median processed signal of the reference channel for probes within the variant was less than 100 when averaged across all arrays (where the processed signal is normalized across each array to have a median value of 1000), or the average reference channel signal for that region was more than 4-fold lower ($\log_2$ ratio < -2) than the average signal for all samples.

GO analysis was conducted using GOstat, which finds statistically over-represented GO terms within a group of genes to generate a list of over-represented GO terms (43). False discovery rate was selected to correct for multiple testing (44).

Multiplex ligation-dependent probe amplification

MLPA was performed to validate a large duplication, in order to determine its breakpoints more precisely and to investigate its Mendelian transmission. Oligonucleotide probe pairs for ligation were designed following the MRC-Holland guidelines for MPLA (Supplementary Material, Table S5) (45). All MLPA reagents were obtained from MRC-Holland, and all reactions were performed as described previously (46). Briefly, DNA samples were heat-denatured in a thermocycler at 98°C for 5 min, and hybridization of the probes was carried out in an overnight incubation at 60°C.

Ligation reactions were then carried out using Ligase-65 mix, and PCR amplification of the ligated probes was performed using a SALSA Polymerase mix, which included universal MLPA primers: 5′ FAM-labelled primer (GGGT TCCCTAAGGGTTGGA) and 3′ primer (TCTAGATTGGA TCTTGCTGGCAC). Two control probes previously validated in another study were included in each experiment (47). MLPA products were separated using an AB 3730x1 DNA Analyser (Applied Biosystems) and outputs were analysed using Gene Mapper software. Data normalization and analysis of peak ratios to determine the copy number of each region were subsequently performed using Microsoft Excel.

PCR-based validation

Validation of 20 deletion polymorphisms was carried out by PCR amplification with paired probe sets spanning and within the deleted region. Primers were designed using Primer3 software (48) (sequences available on request). PCR was carried out using standard methods and products were examined by agarose gel electrophoresis.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

FUNDING

This work was funded by the Hammersmith Hospital NHS Trust (HHNT) Award and the Imperial College Research Excellence Award.

Conflict of Interest statement. None declared.

REFERENCES

1. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. and Pritchard, J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
2. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A. and Cox, D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
3. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
4. Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M. and Eichler, E.E. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, **79**, 275–290.
5. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J. and Altshuler,

D.M. (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
6. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
7. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
8. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R. *et al.* (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.
9. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.
10. Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E. and Lam, W.L. (2007) A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.*, **80**, 91–104.
11. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J. *et al.* (2005) The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.
12. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E. *et al.* (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439**, 851–855.
13. The Autism Genome Project Consortium (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.*, **39**, 319–328.
14. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of *de novo* copy number mutations with autism. *Science*, **316**, 445–449.
15. Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.*, **14**, 417–422.
16. Database of genomic variants—a curated catalogue of structural variation in the human genome. http://projects.tcag.ca/variation/ (March 2007).
17. Carter, N.P. (2004) As normal as normal can be? *Nat. Genet.*, **36**, 931–932.
18. Shianna, K.V. and Willard, H.F. (2006) Human genomics: in search of normality. *Nature*, **444**, 428–429.
19. Risin, S., Hopwood, V.L. and Pathak, S. (1992) Trisomy 12 in Epstein–Barr virus-transformed lymphoblastoid cell lines of normal individuals and patients with nonhematologic malignancies. *Cancer Genet. Cytogenet.*, **60**, 164–169.
20. Lipson, D., Aumann, Y., Ben-Dor, A., Linial, N. and Yakhini, Z. (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput. Biol.*, **13**, 215–228.
21. Ben-Dor, A., Lipson, D., Tsalenko, A., Reimers, M., Baumbusch, L., Barrett, M., Weinstein, J., Børresen-Dale, A.-L. and Yakhini, Z. (2007) Framework for identifying common aberrations in DNA copy number data. *Research in Computational Molecular Biology, Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, **Vol. 4453**.
22. Lipson, D., Tsalenko, A., Yakhini, Z. and Ben-Dor, A. (2005) Interval scores for Quality Annotated CGH DataGENSIPS. http://users.isr.ist.utl.pt/~jmrs/research/Genomics/gensips2005/papers/Gensips2005_SPSAppraoch140.pdf.
23. Fanciulli, M., Norsworthy, P.J., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J.M., Gough, S.C.L., de Smith, A., Blakemore, A.I.F. *et al.* (2007) *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.*, **39**, 721–723.
24. Online Mendelian Inheritance in Man OMIM (TM). McKusick–Nathans Institute of Genetic Medicine Johns Hopkins University (Baltimore MD) National Center for Biotechnology Information National Library of Medicine (Bethesda MD). (2007). http://www.ncbi.nlm.nih.gov/omim/.
25. Rumsey, S.C., Obunike, J.C., Arad, Y., Deckelbaum, R.J. and Goldberg, I.J. (1992) Lipoprotein lipase-mediated uptake and degradation of low density lipoproteins by fibroblasts and macrophages. *J. Clin. Invest.*, **90**, 1504–1512.
26. Watanabe, K., Sugawara, C., Ono, A., Fukuzumi, Y., Itakura, S., Yamazaki, M., Tashiro, H., Osoegawa, K., Soeda, E. and Nomura, T. (1998) Mapping of a novel human carbonyl reductase, *CBR3*, and ribosomal pseudogenes to human chromosome 21q22.2. *Genomics*, **52**, 95–100.
27. Apergis, G.A., Crawford, N., Ghosh, D., Steppan, C.M., Vorachek, W.R., Wen, P. and Locker, J. (1998) A novel nk-2-related transcription factor associated with human fetal liver and hepatocellular carcinoma. *J. Biol. Chem.*, **30**, 2917–2925.
28. Hamilton, A.T., Huntley, S., Tran-Gyamfi, M., Baggott, D.M., Gordon, L. and Stubbs, L. (2006) Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. *Genome Res.*, **16**, 584–594.
29. Ingman, M., Kaesmann, H., Paabo, S. and Gyllensten, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708–713.
30. Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S., Saha, N., Jenkins, T., Tahir, M.A., Deininger, P.L. and Batzer, M.A. (1997) Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.*, **7**, 1061–1071.
31. Locke, D.P., Segraves, R., Carbone, L., Archidiacono, N., Albertson, D.G., Pinkel, D. and Eichler, E.E. (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. *Genome Res.*, **13**, 347–357.
32. Siniscalco, M., Robledo, R., Orru, S., Contu, L., Yadav, P., Ren, Q., Lai, H. and Roe, B. (2000) A plea to search for deletion polymorphism through scans in populations. *Trends Genet.*, **16**, 435–437.
33. Voight, B.F., Kudaravalli, S., Wen, X. and Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, 446–458.
34. Nguyen, D., Webber, C. and Ponting, C.P. (2006) Bias of selection on human copy-number variants. *PLoS Genet.*, **2**, 198–206.
35. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G. *et al.* (2005) A common inversion under selection in Europeans. *Nat. Genet.*, **37**, 129–137.
36. Eichelbaum, M., Ingelman-Sundberg, M. and Evans, W.E. (2006) Pharmacogenomics and individualized drug therapy. *Annu. Rev. Med.*, **57**, 119–137.
37. Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R. and Scherer, S.W. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.*, **1**, e56.
38. Koolen, D.A., Vissers, L.E.L.M., Pfundt, R., de Leeuw, N., Knight, S.J.L., Regan, R., Kooy, R.F., Reyniers, E., Romano, C., Fichera, M. *et al.* (2006) A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.*, **38**, 999–1001.
39. Scherer, S.W., Lee, C., Birney, E., Altshuler, D.M., Eichler, E.E., Carter, N.P., Hurles, M. and Feuk, L. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
40. http://www.chem.agilent.com/scripts/literaturePDF.asp?iWHID=39980. 2007.
41. http://www.chem.agilent.com/scripts/PDS.asp?lPage=29457. 2007.
42. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
43. Beissbarth, T. and Speed, T. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
44. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, **57**, 289–300.
45. http://www.mlpa.com/pages/support_desing_synthetic_probespag.html. 2007.
46. Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, D.F. and Pals, G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, **30**, e57.
47. Stern, R.F., Roberts, R.G., Mann, K., Yau, S.C., Berg, J. and Ogilvie, C.M. (2004) Multiplex ligation-dependent probe amplification using a completely synthetic probe set. *Biotechniques*, **37**, 399–405.
48. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.