nature genetics

Large multiallelic copy number variations in humans

Robert E Handsaker[1–3], Vanessa Van Doren[1–3], Jennifer R Berman[4], Giulio Genovese[1–3], Seva Kashin[1–3], Linda M Boettger[3] & Steven A McCarroll[1–3]

Thousands of genomic segments appear to be present in widely varying copy numbers in different human genomes. We developed ways to use increasingly abundant whole-genome sequence data to identify the copy numbers, alleles and haplotypes present at most large multiallelic CNVs (mCNVs). We analyzed 849 genomes sequenced by the 1000 Genomes Project to identify most large (>5-kb) mCNVs, including 3,878 duplications, of which 1,356 appear to have 3 or more segregating alleles. We find that mCNVs give rise to most human variation in gene dosage—seven times the combined contribution of deletions and biallelic duplications—and that this variation in gene dosage generates abundant variation in gene expression. We describe 'runaway duplication haplotypes' in which genes, including *HPR* and *ORM1*, have mutated to high copy number on specific haplotypes. We also describe partially successful initial strategies for analyzing mCNVs via imputation and provide an initial data resource to support such analyses.

Human genomes exhibit segmental copy number variation (CNV) at thousands of loci. Rare and *de novo* deletions and duplications, which are often large (hundreds of kilobases in length), are known risk factors in many human diseases[1–6]. In addition, thousands of smaller common deletions and duplications segregate in human populations[7,8], many potentially contributing to complex phenotypes[9–12]. Analysis of CNVs, either via direct molecular analysis (for rare CNVs) or statistical imputation (for common CNVs), is now a routine activity in genetic studies[8,13,14].

Perhaps the most intriguing form of CNV is the one that is today least characterized. Many hundreds of genomic segments (and perhaps far more) seem to have copy numbers that vary over wide ranges and have resisted effective analysis by most molecular methods. These loci exist in more states than can be explained by the segregation of just two structural alleles. We and others have called such loci multiallelic CNVs[7,8] (mCNVs), although the specific alleles that segregate at these loci are unknown.

Cytogenetic analysis of a few mCNVs has identified tandem arrays of a genomic segment[15–19]. Such loci may evolve in copy number via non-allelic homologous recombination (NAHR)[20], with mutation rates substantially higher than for SNPs. The actual frequency at which mCNV loci undergo such mutations is unknown, and the process might involve many structural mutations and the repeated recurrence of structurally similar alleles.

An important genome-wide survey by Conrad *et al.*[7] ascertained many mCNVs using high-density arrays to identify such variation in 40 individuals and then analyzed these CNV regions using targeted arrays in 270 individuals. This data set has been the core scientific resource on common CNVs for many years. Reflecting limitations in array-based methods, however, that study inferred integer copy numbers only in the range of 0–5. A subsequent sequencing-based study by Sudmant *et al.*[21] used early whole-genome sequence data from the 1000 Genomes Project pilot to assess CNV at sites annotated as segmental duplications on the human reference genome; this work suggested that hundreds of such loci exhibit CNV, some with wide dynamic ranges in copy number, but studied CNV as a continuous variable, reflecting the analytical challenge of inferring precise integer copy number states[21]. An important scientific need is to understand mCNVs in the genetic terms used to understand other forms of genetic variation: the alleles that generate variation at a site, the frequencies of such alleles and the haplotypes that such alleles form with other variants.

Here we sought to use emerging whole-genome sequence data to answer the following questions. What is the range of integer copy numbers for large mCNVs, and how common is each copy number level? What copy number alleles give rise to such variation? What combinations of rare and common copy number alleles segregate at each locus? How much do mCNVs affect the expression of the genes they contain? By what structural histories did these loci come to their present diversity? How can such variation be incorporated into the analysis of complex traits?

RESULTS

Computational approach and initial results

High copy numbers have been hard to measure experimentally, especially at a genome-wide scale. Precise molecular quantification is challenging because the ratios in DNA content from person to person at mCNVs (such as 4:3 and 7:6) are within the experimental noise of many approaches. Thus, most experimental measurements of mCNV copy number are continuously distributed. Resolving these measures to accurate determinations of the discrete copy

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [4]Digital Biology Center, Bio-Rad Laboratories, Inc., Pleasanton, California, USA. Correspondence should be addressed to S.A.M. (mccarroll@genetics.med.harvard.edu).

Received 28 July 2014; accepted 31 December 2014; published online 26 January 2015; doi:10.1038/ng.3200

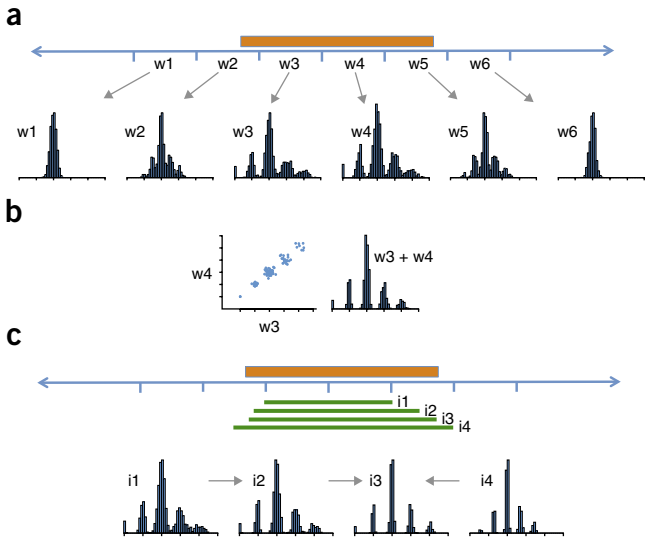

Figure 1 Ascertainment of multiallelic copy number variations (mCNVs) across the human genome. (**a**) Multimodal patterns of variation for a high-frequency CNV (the orange box represents the true extent of the CNV) can be detected in multiple windows (w1–w6) that overlap the CNV segment. (**b**) Where read depth distributions from adjacent windows are highly correlated across many genomes, these windows are merged to increase power for genotyping. (**c**) To more precisely estimate the genome sequence affected, many candidate intervals (green bars; i1–i4) are tested; intervals for which the data most strongly coalesce to integer genotypes with high posterior likelihoods define the estimated CNV boundaries (i3).

number state in each genome is a necessary first step toward a deeper population genetic understanding of mCNVs.

In whole-genome sequence data, the number of sequence reads arising from a genomic segment can reflect the underlying copy number of that segment[21–25]. However, a key challenge is to neutralize the many technical influences that both vary between specific DNA samples or sequencing libraries and reflect the sequence-specific properties of a genomic locus. For example, the GC content of genomic sequences affects their representation in sequencing libraries, owing to PCR amplification bias, in a library-specific manner[21] (**Supplementary Fig. 1**). In DNA samples from proliferating cell lines, such as those used in the 1000 Genomes Project, locus-specific replication timing also influences the read depth of coverage[26]. We found that analyzing many genomes together in a population-based approach[14] could address these and other technical influences (**Fig. 1** and Online Methods).

To obtain precise integer measurements of diploid copy number for mCNVs, we extended the Genome Structure in Populations (Genome STRiP) algorithm[14] to apply its population-based analysis to carefully normalized sequence representation measurements (Online Methods). Within Genome STRiP, we analyzed the distribution of read depth for a genomic segment across many genomes using constrained Gaussian mixture models (**Figs. 1** and **2**, and Online Methods). These models simultaneously infer the most likely integer copy number for each genome and the confidence of each potential integer copy number assignment, which we represent as the 'copy number likelihoods' (the likelihood of each potential copy number 'genotype' given the sequencing data, analogous to genotype likelihoods for SNPs[27–29]). Population-level analysis also enabled us to more accurately infer the correct range of absolute copy numbers by exploiting information inherent in the read depth ratios between different copy number classes and the relative frequencies of different copy number classes (under Hardy-Weinberg equilibrium models) (Online Methods).

The accuracy of this approach at individual loci encouraged us to use it to search the genome *ab initio* for CNVs (in contrast to looking at sites of known segmental duplication[21] or sites with other evidence of duplication, such as aberrantly oriented read pairs[30]). We scanned the human genome in overlapping windows, looking for segments where the read depth measurements deviated from a unimodal distribution (**Fig. 1** and Online Methods). When candidate CNV regions were found, we mapped them at higher resolution by testing multiple segments until the population-level copy number distributions converged to multimodal distributions in which the estimated copy numbers for individuals clustered at a series of integer levels (**Fig. 1** and Online Methods). For CNV loci that were represented on the reference genome by two or more nearly identical segments, we measured total copy number by integrating read depth measurements from positions that were unique across the genome and positions that were identical between the duplicated segments on the reference genome (Online Methods). Such population-scale approaches became more powerful in the simultaneous analysis of many genomes (**Figs. 1** and **2**).

Using these approaches, we generated a detailed ploidy map of 849 genomes from 14 diverse human populations (**Supplementary Table 1**) from Phase 1 of the 1000 Genomes Project, sequenced at a median coverage of 4.8× (range of 2.0–20.6×). We focused on CNV loci where we could obtain a clear series of discrete copy number classes (**Fig. 2**) at this sequencing depth. We ascertained 8,659 CNVs (median length of 10.7 kb), of which 1,449 (16%) overlapped protein-coding genes. Some 4,781 of these CNVs appeared to be biallelic deletions, 2,522 were biallelic duplications and 1,356 were multiallelic (**Supplementary Note**). The distribution of observed diploid copy numbers at these loci ranged from 0 to 15 (**Fig. 2** and **Supplementary Table 2**).

Critical molecular evaluation of CNV and genotype calls

We sought to evaluate, by independent methods, the accuracy of these copy number determinations. Using data from Illumina Omni 2.5 and Affymetrix 6.0 SNP arrays for the same individuals, we applied an intensity rank-sum (IRS) test that analyzed the distributions of array probe intensity measurements across samples to estimate a false discovery rate (FDR) for the identification of CNV loci (Online Methods). This IRS test previously estimated FDRs of 2–83% for various deletion discovery algorithms used in the 1000 Genomes Project[31]. For the CNV discovery set in the current work, this same test estimated an FDR of 2.7%, including 5.8% for duplication CNVs overall and 1.7% for duplication CNVs larger than 10 kb (**Supplementary Table 3**).

We next evaluated the specific integer copy number determinations in each genome (often called genotypes in analogy to crisp SNP genotyping[7,8,32]). Notably, our goal was not just rough correlation of different copy number estimates but determination of the correctness of each integer genotype call in each individual.

For a subset of the CNV sites and individuals, we could compare our results to the high-quality array-based analysis by Conrad *et al.*[7] (**Supplementary Fig. 2**). Across 995 of these CNVs with at least 80% overlap in genome coordinates, our integer genotypes showed 99.9% concordance (exact agreement) with the integer genotype determinations from Conrad *et al.* and agreed at 99.0% of calls of the non-modal copy number (**Supplementary Fig. 3**, **Supplementary Tables 4** and **5**, and **Supplementary Note**). This is the first sequencing-based duplication call set, to our knowledge, to show such high, systematic

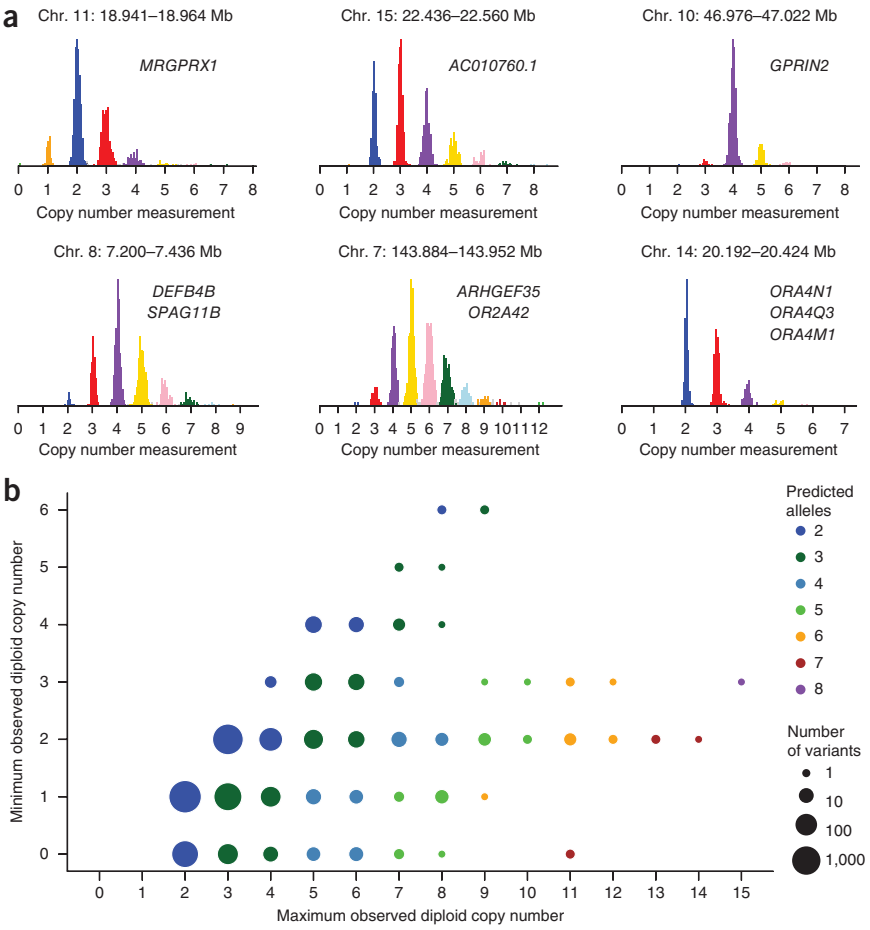

Figure 2 Determination of the copy number levels and alleles present at mCNV loci. (**a**) Histograms of normalized read depth are fitted with a Gaussian mixture model to infer the integer copy number level (genotype) for 849 genomes. Colors represent copy number calls at 95% confidence; samples in gray have less confident copy number calls. The heights of the vertical bars are proportional to the number of genomes with that copy number measurement. Similar plots for all mCNVs ascertained in this study are provided in the accompanying web resource (see URLs). (**b**) Distribution of observed diploid copy numbers across the 8,659 CNVs ascertained in this study. The size of each circle represents the number of CNVs in each category. Colors indicate the minimum number of copy number alleles necessary to generate the observed dynamic range of CNV observed at each site. For example, the blue circles represent deletions and duplications, and the various other colors represent classes of mCNVs with various copy number ranges and numbers of alleles.

genotype concordance with an array-based call set, indicating high accuracy and precision in both data sets.

These validation results were limited to the kinds of CNVs genotyped in the Conrad *et al.* study, where copy number ranged from 0 to 5. To evaluate our genotypes for CNVs with higher copy numbers, we turned to a recently developed molecular method, droplet digital PCR (ddPCR), which uses nanoliter-sized droplets to digitally count the number of copies of a genomic sequence in a DNA sample[33,34]. We selected 22 high-copy CNVs with a wide dynamic range of copy numbers ranging from 1 to 9 (median of 4; mean of 4.47) and typed them using ddPCR in 90 HapMap samples. Integer genotype concordance was 99.9% (**Supplementary Table 6**) and was consistent across the range of copy numbers evaluated (**Fig. 3** and **Supplementary Fig. 4**). These data represent the first such resource, to our knowledge, for high-copy mCNVs (see URLs).

Alleles, phasing and haplotypes

Using these integer copy number determinations in 849 individuals, we next sought to infer the copy numbers present on specific alleles or chromosomal copies. At mCNVs, the diploid copy number of an individual can arise from multiple potential combinations of copy number alleles. (For example, an individual with 7 copies of a genomic locus might have allelic copy numbers of 4 and 3, of 5 and 2, of 6 and 1, or even of 7 and 0.) Additional information exists at a population level, however, as the distribution of copy numbers constrains the relative frequencies of different copy number alleles[35,36]. There is also information in flanking SNP haplotypes that could potentially inform the analysis of individuals for whom multiple allelic combinations are possible[35,36].

For each CNV, we partitioned the diploid copy number likelihoods among all potential combinations of copy number alleles and integrated this information with the genotype likelihoods for flanking SNPs in a population framework, using the Beagle 4 imputation software[37] to phase each CNV (Online Methods). This approach allowed an initial estimate of the frequency of each underlying copy number allele in multiple human populations (see URLs and **Supplementary Fig. 5**).

This analysis identified a wide range of allelic architectures and copy number ranges at mCNVs (**Fig. 2b**). Some 1,356 of the CNVs ascertained in this study were confirmed by this analysis to have 3 or more segregating alleles; 121 of these had 4 or more alleles and 45 of these had 5 or more alleles. (Note that an 'allele' in this analysis refers to the integer copy number of a genomic segment on a single chromosome; the same copy number 'allele' could in principle have arisen multiple times and encompass fine-scale differences that are beyond the scope of the current analysis.) One result of this analysis was that not only the number but also the fraction of duplication CNVs that were multiallelic (35%) was greater than in estimates from earlier studies. These increases are due to the larger sample size with its increased ability to ascertain low-frequency alleles; many duplication CNVs that we and others[7,8] previously observed in only 2 to 4 copies per diploid genome turned out to have additional, higher-copy alleles at low frequency or in other human populations.

mCNVs are the largest source of gene dosage variation in humans

These data made it possible to measure the impact of distinct classes of CNV on human variation in gene dosage. For each human gene, we estimated the frequency with which any pair of individuals differed in the integer copy number of the entire gene, using the copy number determinations from this study (**Supplementary Note**).

Intriguingly, this analysis indicated that mCNVs (a relatively small subset of all CNVs) contributed 88% of the variation in human gene dosage (66 of the 75 gene copy number differences present on average between any pair of individuals), approximately 7 times more

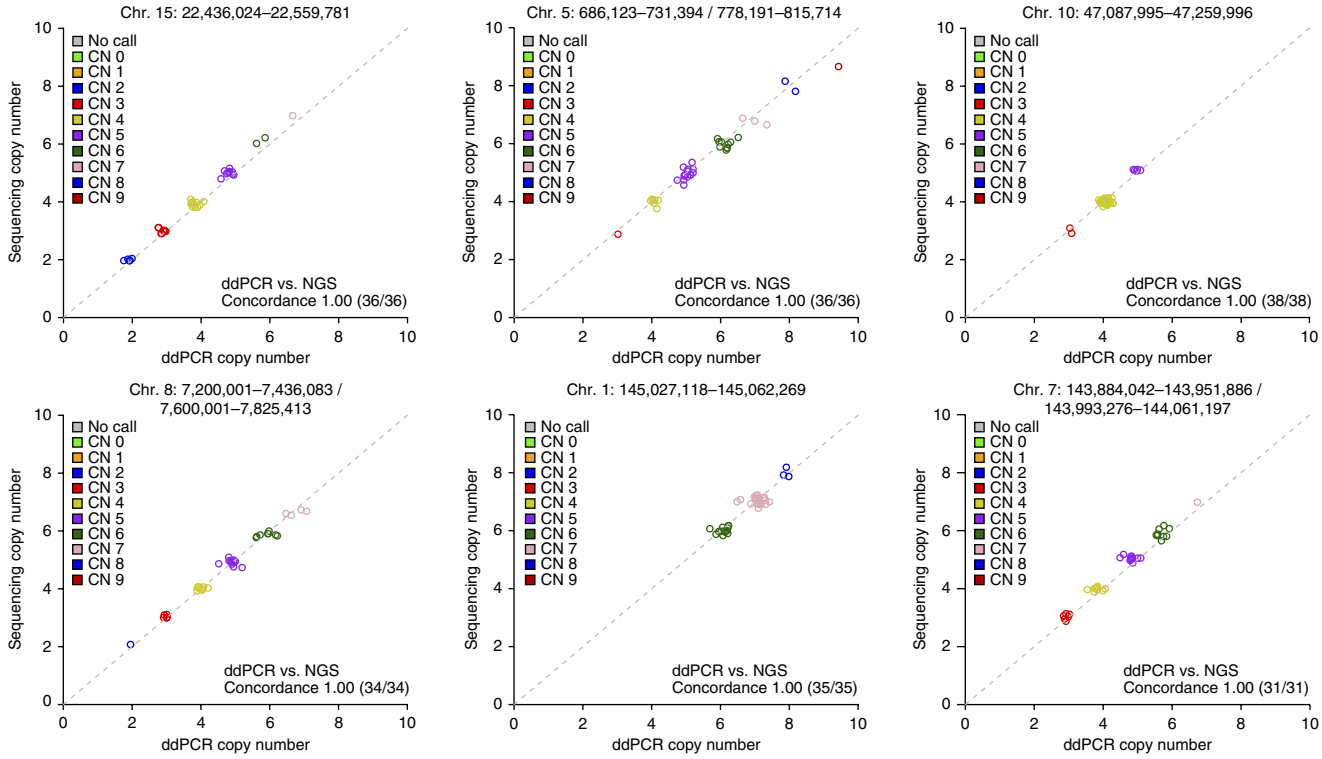

Figure 3 Critical evaluation of copy number genotypes by ddPCR. Across the 38 genomes evaluated, copy number genotypes from sequencing data (NGS) were compared with measurements from ddPCR. Shown here are data for 6 of the 22 loci evaluated. Plots for all loci are shown in **Supplementary Figure 4**. Across the 22 loci, the methods showed 99.9% genotype concordance at confidently called sites.

than the contribution of the more numerous biallelic CNVs (**Table 1**). Several factors accounted for the large contribution of mCNVs to variation in human gene dosage. First, duplications were more likely than deletions to affect protein-coding genes (**Table 1**), as reported previously[7]. Second, mCNVs contributed much more to gene dosage variation than did the more numerous duplication CNVs (**Table 1**). This latter result is largely due to allele frequency; duplication CNVs appear to have a propensity to become multiallelic as they reach higher allele frequencies in populations (**Table 1**, **Supplementary Fig. 6** and **Supplementary Tables 7** and **8**). We hypothesize that this is owing to an increased opportunity for multicopy alleles to encounter one another in diploid genomes, where they may undergo further duplication due to NAHR.

Our data may well underestimate the contribution of mCNVs to variation in gene dosage, as we have focused on mCNVs for which we could infer highly accurate integer genotypes, limiting the current analysis to CNVs with copy numbers generally less than 12. CNVs with even higher copy numbers appear to exist in human populations[21] and are likely to be multiallelic.

Table 1 CNV impact on gene dosage

				Genes differing in copy number between two individuals				
Category	CNVs	Genic CNVs	Genes	Singletons	AAF < 1%	AAF < 5%	AAF > 5%	Overall
Deletion	4,781	70	88	0.12	0.33	0.54	5.04	5.58
Duplication	2,522	194	314	0.40	1.28	2.32	1.06	3.38
Multiallelic	1,356	126	231	NA	0.80	2.62	63.27	65.89
Total	8,659	390	633	0.52	2.41	5.48	69.37	74.85

Contributions of three forms of CNV (deletions, biallelic duplications and mCNVs) to gene dosage variation among humans. mCNVs, which comprise only about 15% of the CNVs ascertained, give rise to more than 85% of the variation in human gene dosage, measured as the number of genes (on average) that differ in copy number between randomly chosen pairs of individuals. AAF, alternate (non-reference) allele frequency; NA, not applicable.

Biallelic and multiallelic CNVs tended to be over-represented in the same kinds of genes. As reported previously for CNVs in general[7], mCNVs were over-represented among genes with roles in extracellular biological processes and were under-represented among genes involved in intracellular metabolic and biosynthetic pathways (**Supplementary Fig. 7** and **Supplementary Table 9**).

Impact on gene expression

Given the large contribution of mCNVs to human gene dosage variation, we sought to understand their contribution to variation in gene expression. In the simplest model, increased dosage of a gene could directly cause increases in the expression of that gene. However, several factors might modify or abrogate such relationships; for example, tandem repeats could lead to dosage-dependent inactivation, as observed for transgenes in plants, flies and mammals[38–41], or gene duplications might not include distal regulatory elements important for the expression of the gene.

To address this question in one cell type, we used available mRNA sequencing (mRNA-seq) data derived from lymphoblastoid cell lines (LCLs) for 310 individuals whose genomes we analyzed here[42] and asked whether variation in gene dosage was reflected in mRNA abundance. Most genes mapping to mCNVs showed a strongly positive correlation between the dosage of the gene in genomic DNA and the mRNA expression of the gene (**Fig. 4** and **Supplementary Fig. 8**). RNA abundance tended to increase proportionally with gene dosage (**Fig. 4**). Among the 133 genes with evaluable expression levels in LCLs (reads per kilobase of transcript per

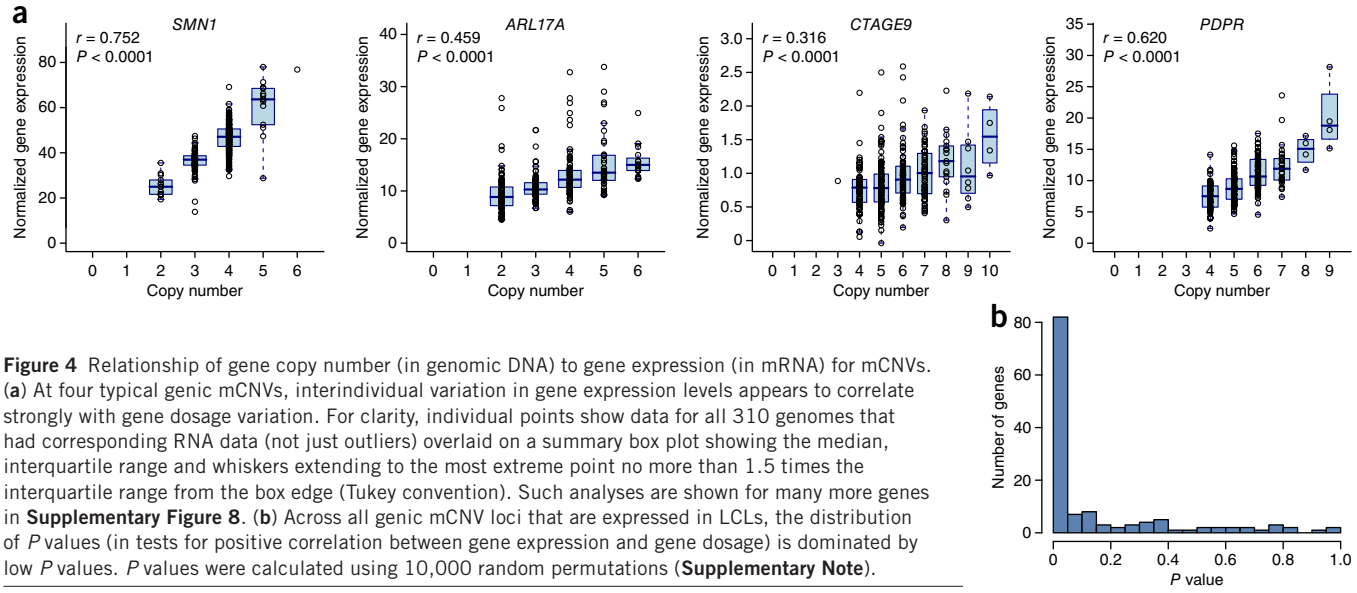

Figure 4 Relationship of gene copy number (in genomic DNA) to gene expression (in mRNA) for mCNVs. (**a**) At four typical genic mCNVs, interindividual variation in gene expression levels appears to correlate strongly with gene dosage variation. For clarity, individual points show data for all 310 genomes that had corresponding RNA data (not just outliers) overlaid on a summary box plot showing the median, interquartile range and whiskers extending to the most extreme point no more than 1.5 times the interquartile range from the box edge (Tukey convention). Such analyses are shown for many more genes in **Supplementary Figure 8**. (**b**) Across all genic mCNV loci that are expressed in LCLs, the distribution of *P* values (in tests for positive correlation between gene expression and gene dosage) is dominated by low *P* values. *P* values were calculated using 10,000 random permutations (**Supplementary Note**).

million mapped reads (RPKM) > 2) that were contained within a common duplication CNV, the distribution of association *P* values was dominated by low *P* values, indicating a predominance of positive correlations, with only a few potential exceptions (**Fig. 4**). We conclude that the great majority of mCNVs affect the RNA expression levels of the genes they contain. We further explore the relationships between mCNVs and gene expression in **Supplementary Figure 9** and **Supplementary Table 10**.

Imputation of the allelic states of mCNVs from SNPs

The above results show that hundreds of human genes exhibit common variation in dosage, structure and expression due to mCNVs. It will be important to understand how such variation contributes to human phenotypes. An initial study of common CNVs in 6 common diseases (1,500 cases per disease) by the Wellcome Trust Case Control Consortium found few effects from common CNVs[43]; however, most discoveries of SNP associations for complex phenotypes have required much larger samples (tens of thousands) and have been made during meta-analysis of many genome-wide association study (GWAS) data sets. Similar discoveries might be enabled if the states of mCNVs could be imputed from available SNP data. We recently found that the 17q21.31 locus segregates in human populations in at least nine structural forms[34], of which most are accessible to imputation[34]. However, the generality of such relationships is not known.

We sought to understand the extent to which mCNV imputation might be possible using our results as an imputation resource. Using the Beagle 4 software[37] to perform imputation, we evaluated how effectively diploid copy number could be estimated from flanking SNP data for each CNV by performing leave-some-out analyses (Online Methods).

It is helpful to compare the results for mCNVs to those for simple deletion CNVs. Common deletion CNVs (a positive control) were overwhelmingly well imputed (91% with $r^2 > 0.8$), as expected from earlier work. By contrast, mCNVs showed a wide range of imputability, with common mCNVs (combined minor allele frequency (MAF) > 10%) exhibiting an almost uniform distribution of imputation r^2 values from 0 to 1.0 (**Supplementary Fig. 10**). These results suggest that mCNVs have a wide range of imputability, from mCNVs at which each copy number allele is well imputed to mCNVs with frequently recurring mutations and little sustained relationship of copy number to surrounding haplotypes.

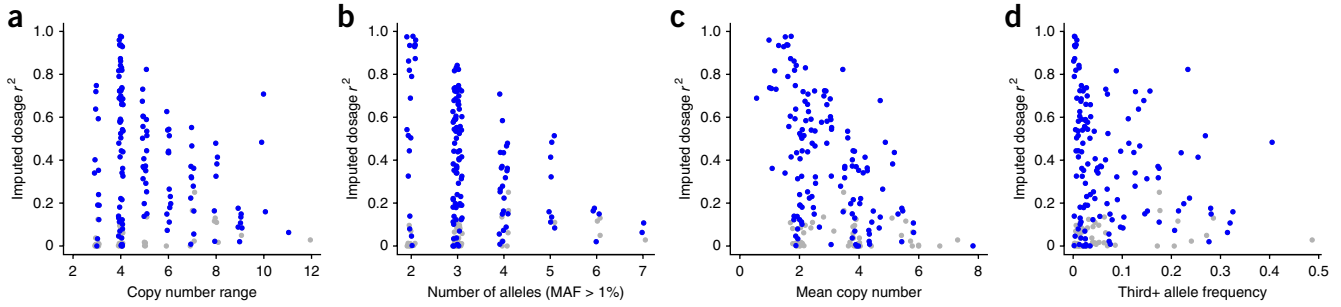

Figure 5 Relationship between the imputability of mCNVs and the features of each mCNV locus. (**a**–**d**) Imputability from current reference panels (here measured by imputed dosage r^2) relates to multiple features of each mCNV, including the copy number range of the mCNV (the difference between the highest and lowest observed diploid copy numbers) (**a**), the number of common (MAF > 1%) copy number alleles segregating at the site (**b**), the mean copy number of the mCNV (**c**) and the combined frequency of all copy number alleles excluding the two most common ones (third+ allele frequency) (**d**). All quantities were calculated for the EUR (European) population cohort for 184 mCNVs with MAF > 1%. In **a** and **b**, a small amount of random variation was added to the discrete *x*-axis values to aid visualization. CNVs for which at least one individual SNP showed even partial correlation with copy number ($P < 1 \times 10^{-3}$) in the EUR population are plotted in blue; CNVs lacking such SNPs are plotted in gray.

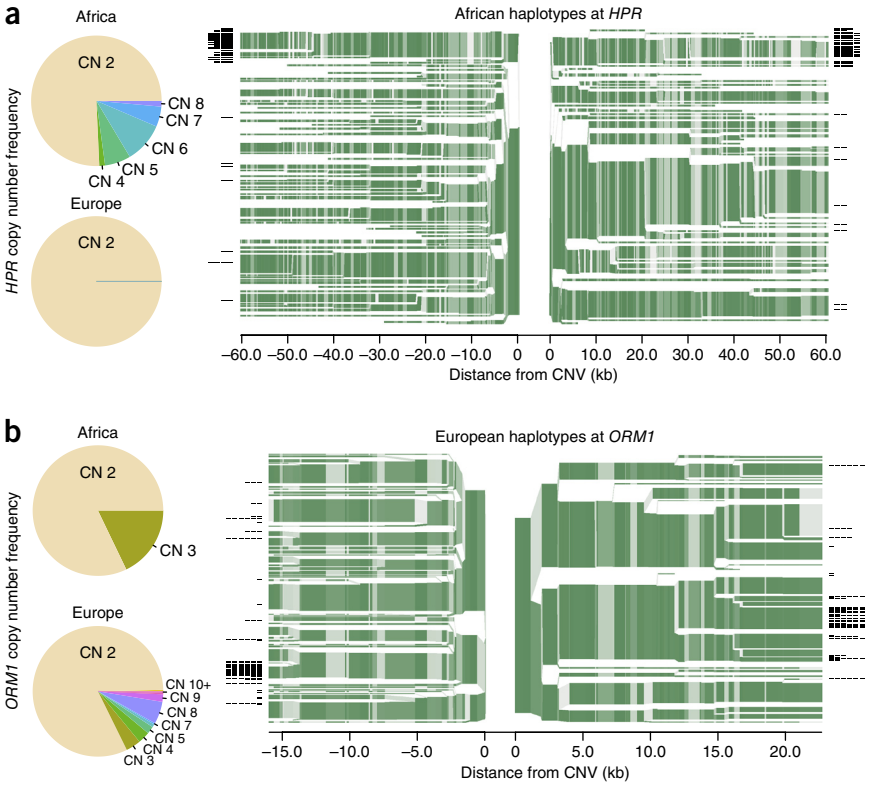

Figure 6 Haplotypes with runaway copy number. (**a**) Copy number distribution and haplotype structure of an mCNV encompassing the *HPR* gene. About 25% of the non-admixed African individuals sampled by the 1000 Genomes Project exhibit *HPR* copy numbers greatly increased (4–8) relative to those observed in individuals sampled from all the non-African populations (generally no more than 2). The branching green plots on the right show SNP haplotypes in the region around the *HPR* locus for chromosomes sampled from African populations (YRI (Yoruba from Ibadan, Nigeria) and LWK (Luhya in Webuye, Kenya)). The origin in the middle of the haplotype plot corresponds to the edges of the *HPR* mCNV; the branches show places at which flanking haplotypes begin to diverge because of mutation or recombination. The thickness of each branch indicates haplotype frequency; shading indicates the allele frequency of the individual SNPs used to define haplotypes. Haplotypes carrying high-copy *HPR* alleles (with more than 1 *HPR* copy) are indicated by black lines at branch tips, with a line segment for each extra copy above 1. Almost all the high-copy alleles appear to segregate on the same haplotype background. (**b**) A similar analysis of an mCNV affecting the *ORM1* gene, which appears to have greatly expanded in copy number on a specific haplotype, producing many different high-copy alleles.

A scenario that could potentially hinder the imputation of some mCNVs would occur if the duplication alleles were dispersed to distant genomic sites[44,45]. We looked for evidence of dispersed duplications using long-range linkage disequilibrium (LD) and interchromosomal segmental duplications and found 15 CNVs with evidence of dispersal (**Supplementary Fig. 11** and **Supplementary Tables 11** and **12**); however, these were a small fraction (2.2%) of the duplications evaluated, suggesting that distant dispersal is unlikely to explain the modest efficacy of imputation for many mCNVs.

Many features of the allelic architecture and copy number distributions of mCNVs were predictive of their imputability. Lower average imputability was evident for mCNVs with wider copy number ranges (**Fig. 5a**), larger numbers of alleles (**Fig. 5b**), higher average copy number (**Fig. 5c**) or a high 'third+ allele frequency' (when the less common alleles of an mCNV had relatively high frequencies) (**Fig. 5d**). All of these features of mCNVs are likely to be proxies for their historical mutation rates, the complexities of their mutational histories and the consequent likelihood that a given flanking SNP haplotype presents with different copy number states in different individuals.

Runaway duplication haplotypes

Some genic CNVs showed an intriguing pattern in which most individuals had low copy numbers, but some had far higher copy numbers; the high copy numbers appeared to arise from alleles on which a genomic segment had duplicated many times. In each case, the individuals with high copy numbers were from the same continental population. One of the genes with this type of CNV was *HPR*, which encodes a haptoglobin-related protein that is used in defense against trypanosomes[46]. *HPR* was present at 2 copies in European population samples and at 1–2 copies in Asian population samples, but it was present at 4–8 copies in about 25% of individuals sampled from African populations (**Fig. 6**). A second example was the *ORM1* gene (encoding orosomucoid), which was present at 2–3 copies (per diploid genome) in all African genomes analyzed but at up to 13 copies among Europeans, with high copy numbers particularly common among southern Europeans (**Fig. 6**).

To better understand the mutational history of each locus, we analyzed the haplotypes on which the low- and high-copy alleles were segregating. At both *ORM1* and *HPR*, the high-copy alleles were generally observed on a shared SNP haplotype, whereas the reference structural allele (with a single gene copy) segregated on many different haplotypes (**Fig. 6**). The haplotype uniformity and population specificity of the high-copy alleles suggest a recent, geographically unique origin. These relationships also indicate that the high-copy alleles have evolved by repeated, recurrent mutation on the same haplotype background.

Both *HPR* and *ORM1* are functionally connected to disease-associated loci. The HPR protein forms (together with APOL1) the trypanosome lytic factor (TLF) that is a primary defense of human blood against trypanosomes[46,47]. *APOL1* is hypothesized to have been under strong recent selection in African populations due to the presence of recently arisen variants that protect against trypanosome infection while simultaneously contributing to kidney disease[48]. The CNV at *HPR* is similar to the African-specific SNPs at *APOL1* in that it contains multiple new alleles that have quickly risen to high frequencies in African populations in which trypanosomes are endemic[48,49]. *ORM1*, which encodes one of the most abundant glycoproteins in blood, is paralogous to *ORMDL3*, in which common variants associate strongly with risk of asthma[50]. It is intriguing to speculate that *HPR* and/or *ORM1* might have quickly increased in copy number in response to geographically localized selection events.

Both *HPR* and *ORM1* present examples of the partial efficacy of imputation for analysis of mCNVs. At both loci, SNP haplotypes readily distinguished the high-copy alleles from the more common

low-copy alleles (**Fig. 6**). However, SNP haplotypes did not readily distinguish the high-copy alleles from one another (**Fig. 6**). It seems likely that high-copy alleles have frequently mutated into other high-copy states, whereas low-copy alleles have remained stable. Understanding the actual mutation rates of mCNVs is an important direction for future work.

DISCUSSION

In this study, we have described computational approaches for identifying mCNVs and characterizing such polymorphisms in terms of alleles, allele frequencies and haplotypes. The resulting data show that mCNVs broadly affect genes and gene expression. mCNVs give rise to at least six times more variation in gene dosage than do simpler, biallelic CNVs (**Table 1**), and such variation usually affects the expression levels of the encompassed genes (**Fig. 4**).

mCNVs are not routinely evaluated in genome-wide studies based on SNP arrays or exome sequencing because of the limitations of these approaches in accurately measuring copy numbers greater than 4. Two studies have analyzed mCNVs using custom-designed targeted arrays, which appear to be more effective, although both studies noted the immense technical challenges involved and the consequent challenges of careful association analysis when copy number estimates are not precise integer genotypes[43,51]. Several features of mCNVs suggest that they will reward future efforts at analysis. First, most mCNVs affect just one or two genes, allowing functional effects to be assigned to specific genes. Second, associations with mCNVs will exhibit a clear direction of effect (with certainty about whether more or less gene activity contributes to risk). Third, therapeutics based on such discoveries might benefit from the observation that dosage variation is present and tolerated in the general population.

But what is the best way to relate mCNVs to phenotypes? Abundant whole-genome sequencing data, together with analysis methods such as those described here, could eventually make direct association testing routine. But it will take many years (and substantial resources) for the sample size of whole-genome sequencing–based studies to approach the current size of array-based GWAS. For now, we believe that early insights might be gleaned from new analyses of large extant SNP data sets.

Imputation from existing SNP data could be used to perform initial genome-wide scans to nominate specific mCNV loci for deeper analysis. Given the finding here that an imputation-derived dosage estimate will only partially correlate with true copy number, such an approach would be only partially powered but could draw compensating power from the vast numbers of individuals for whom dense SNP data are available. On the basis of observing at least nominal association of a variant with a phenotype, follow-up evaluation could involve direct molecular analysis (which might strengthen the association signal) and conditional analysis relative to nearby variants. The finding that biallelic SNPs only partially correlate with mCNV copy number at many loci could empower conditional association analysis to disentangle the effects of mCNV alleles from those of other nearby genetic variants.

There are many important directions for future work. The analysis of mCNVs in larger population-based cohorts will yield new insights about the combinations of common and lower-frequency structural alleles that reside at each locus; such cohorts could also be used to create second-generation imputation resources that are more effective than the one described here at imputing copy numbers for recurrently mutating mCNVs. Our work here has not addressed more complex forms of structural variation in which multiple duplications and deletions (affecting overlapping but distinct genomic segments) and inversions give rise to complex, compound structural alleles; such loci have thus far required customized computational and molecular strategies[34,44,52,53].

The elucidation of complex and recurrently mutating forms of genome variation will ultimately deepen the understanding of genome evolution and the ways in which these mutable loci contribute to human phenotypes.

URLs. The catalog of CNVs ascertained in this study, including genotypes, allele frequencies and imputation results (with plots of flanking haplotypes), is available at http://www.broadinstitute.org/software/genomestrip/mcnv_supplementary_data.

Software. Genome STRiP software can be downloaded from our web site at http://www.broadinstitute.org/software/genomestrip.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Skvortsov, M. Thornton, N. Klitgord and B. Zhang for contributions to ddPCR assay design and validation. We also thank members of the 1000 Genomes Project for helpful conversations about analysis methods. This work was supported by a grant from the National Human Genome Research Institute (NHGRI; R01 HG006855). An additional grant from NHGRI (U01 HG006510) is supporting follow-on work to develop these methods into production-ready software that can be used by any research laboratory.

AUTHOR CONTRIBUTIONS

R.E.H. and S.A.M. designed the study. R.E.H. devised the computational approaches, performed the analysis and wrote the Genome STRiP software. V.V.D. performed the ddPCR experiments and initial data analyses. J.R.B. designed assays for the ddPCR experiments and provided technical guidance and materials. G.G. contributed to the statistical analyses of gene dosage and dispersed duplications. S.K. helped automate and refine the algorithms and produce the public software release. L.M.B. contributed to the analysis of the *HPR* locus. S.A.M. and R.E.H. interpreted the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
2. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
3. Weiss, L.A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
4. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
5. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–670 (2010).
6. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene *VIPR2* confer significant risk for schizophrenia. *Nature* **471**, 499–503 (2011).
7. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
8. McCarroll, S.A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174 (2008).
9. de Cid, R. *et al.* Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
10. McCarroll, S.A. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41**, 1341–1344 (2009).
11. Willer, C.J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).

12. McCarroll, S.A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
13. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
14. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* **43**, 269–276 (2011).
15. Hollox, E.J., Armour, J.A. & Barber, J.C. Extensive normal copy number variation of a β-defensin antimicrobial-gene cluster. *Am. J. Hum. Genet.* **73**, 591–600 (2003).
16. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
17. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* **39**, S48–S54 (2007).
18. Perry, G.H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
19. Perry, G.H. *et al.* The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* **82**, 685–695 (2008).
20. Gu, W., Zhang, F. & Lupski, J.R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1**, 4 (2008).
21. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
22. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
23. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**, 1586–1592 (2009).
24. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
25. Bellos, E., Johnson, M.R. & Coin, L.J. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.* **13**, R120 (2012).
26. Koren, A. *et al.* Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
27. Wang, Y., Lu, J., Yu, J., Gibbs, R.A. & Yu, F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* **23**, 833–842 (2013).
28. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
29. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
30. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
31. Mills, R.E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
32. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
33. Hindson, B.J. *et al.* High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **83**, 8604–8610 (2011).
34. Boettger, L.M., Handsaker, R.E., Zody, M.C. & McCarroll, S.A. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat. Genet.* **44**, 881–885 (2012).
35. Su, S.Y. *et al.* Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics* **26**, 1437–1445 (2010).
36. Kato, M., Nakamura, Y. & Tsunoda, T. An algorithm for inferring complex haplotypes in a region of copy-number variation. *Am. J. Hum. Genet.* **83**, 157–169 (2008).
37. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
38. Assaad, F.F., Tucker, K.L. & Signer, E.R. Epigenetic repeat-induced gene silencing (RIGS) in *Arabidopsis*. *Plant Mol. Biol.* **22**, 1067–1085 (1993).
39. Dorer, D.R. & Henikoff, S. Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila*. *Cell* **77**, 993–1002 (1994).
40. Dorer, D.R. & Henikoff, S. Transgene repeat arrays interact with distant heterochromatin and cause silencing in *cis* and *trans*. *Genetics* **147**, 1181–1190 (1997).
41. Garrick, D., Fiering, S., Martin, D.I. & Whitelaw, E. Repeat-induced gene silencing in mammals. *Nat. Genet.* **18**, 56–59 (1998).
42. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
43. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–720 (2010).
44. Abu Bakar, S., Hollox, E.J. & Armour, J.A. Allelic recombination between distinct genomic locations generates copy number diversity in human β-defensins. *Proc. Natl. Acad. Sci. USA* **106**, 853–858 (2009).
45. Dennis, M.Y. *et al.* Evolution of human-specific neural *SRGAP2* genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
46. Smith, A.B., Esko, J.D. & Hajduk, S.L. Killing of trypanosomes by the human haptoglobin-related protein. *Science* **268**, 284–286 (1995).
47. Harrington, J.M., Howell, S. & Hajduk, S.L. Membrane permeabilization by trypanosome lytic factor, a cytolytic human high density lipoprotein. *J. Biol. Chem.* **284**, 13505–13512 (2009).
48. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
49. Genovese, G., Friedman, D.J. & Pollak, M.R. *APOL1* variants and kidney disease in people of recent African ancestry. *Nat. Rev. Nephrol.* **9**, 240–244 (2013).
50. Moffatt, M.F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
51. Zanda, M. *et al.* A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes. *PLoS Genet.* **10**, e1004367 (2014).
52. Hollox, E.J. *et al.* Psoriasis is associated with increased β-defensin genomic copy number. *Nat. Genet.* **40**, 23–25 (2008).
53. Steinberg, K.M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012).

ONLINE METHODS

CNV discovery and genotyping. Discovery and genotyping of CNVs was performed using an enhanced version of Genome STRiP software[14] (internal version 1.04.1383; public release 2.0). Genome STRiP performs discovery and genotyping of CNVs by analyzing the data from many samples simultaneously in a population-based framework.

CNV site discovery was performed in two phases: one targeting uniquely alignable portions of the reference genome (for this data set, positions where 36-bp reads could be uniquely aligned) and one targeting segmentally duplicated regions where the copy number in the reference genome was greater than 1. For the uniquely alignable portion of the genome, we used a pipeline that prospectively genotyped overlapping windows across the genome, using a window size of 5 kb of alignable sequence and overlapping adjacent windows by 2.5 kb (discovery set 1; **Supplementary Note**). For segmentally duplicated regions, we used the segmental duplication annotations from the UCSC Genome Browser to define potential CNV regions that were then evaluated by prospective genotyping, assuming a reference allele copy number of 2. The identified regions were subsequently filtered to select sites with clear evidence of polymorphism (discovery set 2; **Supplementary Note**). In both cases, site discovery was enhanced by improvements to the genotyping methods in Genome STRiP, particularly the normalization and interpretation of read depth information.

Normalization of read depth. Read depth of coverage was measured by counting the sequenced DNA fragments overlapping a genomic interval of interest. To facilitate normalization, each fragment was counted as a point event, assigned to the coordinate that was the midpoint of the left-most read (for paired-end sequencing). Reads were only counted if they mapped to locations that should be uniquely alignable according to the structure of the reference genome. In addition, for CNV analysis, we applied a 'low-complexity mask' that masked genome coordinates falling in regions of low-complexity sequence (as categorized by the RepeatMasker tracks from the UCSC Genome Browser). Raw read counts were normalized to correct for sequencing bias as a function of the GC content of each sequencing library (**Supplementary Note**).

CNV genotyping model. To genotype a CNV interval, Genome STRiP fits a constrained Gaussian mixture model to the read depth signal across the interval, using data from all available DNA samples. The model incorporates sample-specific variance terms to model the variation in sequencing depth between samples. In our previous work on deletion variants, we used a mixture of three distributions, corresponding to diploid copy number classes of 0–2. For CNV genotyping, we fitted a mixture of multiple Gaussian distributions, corresponding to a series of diploid copy number classes from 0 to a site-specific maximum. The maximum copy number modeled for each site was chosen on the basis of the maximum read depth signal from any of the samples at that site. An advantage of the constrained Gaussian mixture model used in Genome STRiP is that, in practice, the mixture weights can be allowed to go to zero without adversely affecting the model fit. This eliminates the need to test and compare many different models with different numbers of copy number classes.

Assignment of absolute copy number. A key problem for multi-sample CNV calling algorithms that use clustering is to accurately estimate the correct absolute copy number for each cluster. The constrained mixture model used in Genome STRiP is advantageous in this respect, especially when large numbers of genomes are available for simultaneous analysis. The means of the copy number classes are required to scale as integer multiples (with a scaling parameter fitted from the data). Thus, the model is sensitive to the ratio between adjacent clusters, which can help to distinguish, for example, clusters of copy number 2–4 from clusters of copy number 4–6. To avoid overfitting, we rejected models where the scaling parameter was too high (above 2.0) or too low (below 0.5).

Using copy number parity. To further increase the accuracy of absolute copy number determinations, especially at high copy number, we borrowed information from the population allele frequencies of the samples being jointly analyzed. At autosomal loci, under an assumption of Hardy-Weinberg equilibrium, the number of individuals who have even diploid copy numbers should not be less than the number of individuals with odd diploid copy numbers. (Intuitively, this is a generalization to mCNVs of the observation for SNP genotypes that the frequency of a heterozygote class should not exceed the combined frequencies of the homozygote classes.) At most sites, the allele frequencies will fall into a range where incorrectly shifted copy number assignments will cause strong deviation from Hardy-Weinberg equilibrium. For example, a CNV site with low to moderate variant frequency and diploid copy numbers of 2–4 will have a very different inferred allele frequency distribution if the diploid copy numbers are incorrectly shifted down to 1–3 or up to 3–5. Using this information increases the effective separation between likely cluster assignments by a factor of two at most CNV sites.

We exploited this information by performing a simple parity test on copy number. As we fit the mixture model, we estimated the number of even and odd copy numbers observed in the population. If the fraction of even copy numbers fell below a specified threshold parameter (default of 0.4), we shifted the cluster assignments toward a more likely model. This optimization can be disabled for small populations, family studies or highly stratified populations.

Genotyping both unique and duplicated sequences. Previous versions of Genome STRiP analyzed read depth only at positions on the reference genome that are sufficiently unique that sequence reads should align uniquely (on the basis of read length and the repetitive structure of the reference genome). To genotype CNVs that are present in multiple copies on the reference genome, we extended our method to use positions that are non-unique on the reference genome by considering reads in reference *k*-mers that are not globally unique but are present only at a small, fixed number of locations within paralogous genomic regions. The normalized read counts from such positions can be used to estimate the total copy number of a non-unique segment on the reference genome.

In segmentally duplicated CNV regions, the read counts from both the unique and non-unique positions can be used together to estimate both the total copy number across the genome and the paralog-specific copy number. The unique positions represent paralog-specific differences (or paralog-specific variations, PSVs) that differentiate one paralog from the other, on the basis of the reference genome.

CNV boundary determination. To determine the likely boundaries of a detected CNV, we employed a hill-climbing algorithm that tested many candidate CNV segments overlapping the CNV region to find the segment where the read depth distribution converged most strongly with a series of confidently predicted integer copy numbers in all analyzed samples (**Supplementary Note**). Starting with an initial CNV segment, we sampled the space of potential segment boundaries by alternately varying the left and right boundaries of the segment by several multiples of a fixed increment (10% of the initial segment length), then gradually halving the increment until a target boundary precision (200 bp) or minimum segment length (2.5 kb) was reached.

CNV phasing and imputation. Before phasing, the computed likelihoods for each diploid copy number call were converted to genotype likelihoods for haploid copy number alleles. At mCNV loci, the constituent copy number alleles are generally not known (for example, a diploid copy number of 4 can arise from allelic copy numbers of 2 + 2 or 1 + 3). We first inferred the set of likely copy number alleles by enumerating all possible allelic combinations and employing an expectation maximization algorithm to estimate the allele frequencies. We then considered only alleles that met a population-specific allele frequency threshold of 0.001 (**Supplementary Note**). We used these population-specific estimated allele frequencies as a prior on the genotype likelihoods as additional information to help resolve ambiguous allelic combinations (**Supplementary Note**).

Phasing and subsequent imputation experiments were performed using the Beagle software package (Beagle 4, version r1128) and the 1000 Genomes Project Phase 1 reference panel supplied with Beagle software. Each CNV site was phased and imputed separately. For CNVs at segmentally duplicated

sites where the segments were separated by more than 100 kb, we attempted to phase and impute the CNV separately at each genomic location.

Imputation accuracy was evaluated by performing a series of leave-out trials at each CNV site. In each trial, the CNV genotypes for ten individuals were masked and the genotypes for these individuals were imputed from the rest of the cohort. This was repeated for the entire cohort using disjoint sets of ten individuals at a time. The imputed genotypes were compared to the measured diploid copy numbers for the withheld individuals, using the squared correlation (Pearson's r) between the inferred diploid copy number in each individual (**Supplementary Note**).

Estimation of CNV false discovery rate using SNP array data. We used the IRS test implemented in the Genome STRiP software package to evaluate FDR. This test ranks the normalized array probe intensity values for the SNPs underlying each CNV and performs a rank-sum test across the set of array probes to determine whether samples with genotypes that are below (or above) the expected reference copy number have statistically lower (or higher) probe intensities. The FDR was estimated from the distribution of P values across the evaluated sites, considering deletions and duplication sites separately (**Supplementary Note**). The overall FDR for the call set was estimated as a weighted average of the rates for each variant category.

CNV genotype analysis with droplet digital PCR. We screened a subset of 96 genomic DNA samples from the YRI2 cohort of the Coriell 1000 Genomes Project samples for 32 mCNV and segmental duplication target sites ascertained from sequencing data (**Supplementary Table 13**) using ddPCR (Bio-Rad Laboratories) and standard protocols (**Supplementary Note**). Primer-probe ddPCR assays for each target were custom designed by Bio-Rad Laboratories on the basis of a set of target genomic regions from sequencing data analysis. At CNV sites in segmental duplications on the reference genome, the assays were designed to genomic locations with identically duplicated sequence to measure total copy number across the duplicated segments.

Effect of gene dosage on gene expression. RNA sequencing data were downloaded from the Geuvadis Project web site. For CNVs that fully overlapped genes, we computed the Pearson's correlation coefficient between the integer copy number for the CNV and the normalized level of gene expression determined from the RNA sequencing data.

P values for gene expression were calculated using 10,000 random permutations and a one-sided test, with the Pearson's correlation coefficient as the test statistic. In each trial, we randomly permuted the mapping between genes and CNVs. To control for potential interactions between genes and nearby but non-overlapping CNVs, we generated permutations where genes were always assigned to CNVs on a different chromosome.

doi:10.1038/ng.3200