Published in final edited form as:

*Nat Genet*. 2015 March ; 47(3): 296–303. doi:10.1038/ng.3200.

# Large multi-allelic copy number variations in humans

**Robert E. Handsaker**[1,2,3], **Vanessa Van Doren**[1,2,3], **Jennifer R. Berman**[4], **Giulio Genovese**[1,2,3], **Seva Kashin**[1,2,3], **Linda M. Boettger**[3], and **Steven A. McCarroll**[1,2,3]

[1] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[2] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[3] Department of Genetics, Harvard Medical School, Boston, MA, USA

[4] Digital Biology Center, Bio-Rad Laboratories, Inc., Pleasanton, CA, USA

## Abstract

Thousands of genome segments appear to be present in widely varying copy number 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, multi-allelic 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 three or more segregating alleles. We find that mCNVs give rise to most human gene-dosage variation – exceeding sevenfold the 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 describe partially successful initial strategies for analyzing mCNVs via imputation and provide an initial data resource to support such analyses.

## Introduction

Human genomes exhibit segmental copy number variation (CNV) at thousands of loci. Rare and *de novo* deletions and duplications, which are often large (100s of kb), are known risk factors in many human diseases[1-6]; thousands of smaller, common deletions and

Please address correspondence to mccarroll@genetics.med.harvard.edu..

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. 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

J.R.B. is an employee of Bio-Rad Laboratories, Inc.

Software Availability

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

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 form that is today least characterized. Many hundreds of genomic segments (and perhaps far more) seem to vary in copy number in 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 “multi-allelic CNVs” (mCNVs)[7,15], though the specific alleles that segregate at these loci are unknown.

Cytogenetic analysis of a few multi-allelic CNVs has revealed tandem arrays of a genomic segment[16-20]. Such loci may evolve in copy number via non-allelic homologous recombination (NAHR)[21], with mutation rates substantially higher than for SNPs. The actual frequency with which mCNV loci undergo such mutations is unknown, and might involve many structural mutations and the repeated recurrence of structurally similar alleles.

An important genome-wide survey of CNV by Conrad et al.[7] ascertained many mCNVs using high-density arrays to ascertain CNV in 40 individuals, 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, the Conrad study inferred integer copy numbers only in the range of 0-5. A subsequent sequencing-based study by Sudmant *et al.* used early whole-genome sequence data from the 1000 Genomes Project pilot to assess CNV at sites annotated as segmental duplications on the human genome reference[22]; this work suggested that hundreds of such loci exhibit CNV, some with wide dynamic range, but studied CNV as a continuous variable, reflecting the analytical challenge of inferring precise integer copy-number states[22]. 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 these questions: What is the range of integer copy number 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 genome scale. Precise molecular quantitation 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 to accurate determinations of the discrete copy number state in each genome is a necessary first step towards 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 [22-26.] However, a key challenge is to neutralize the many technical influences that both (i) vary between specific DNA samples or sequencing libraries, and (ii) also reflect sequence-specific properties of a genomic locus. For example, the G+C content of genomic sequences affects their representation in sequencing libraries due to PCR amplification bias, in a library-specific manner[22] (**Supplementary Figure 1**). In DNA samples from proliferating cell lines, such as those used in the 1000 Genomes Project, locus-specific replication timing also influences read depth of coverage[27]. We found that analyzing many genomes together in a population-based approach[14] can address these and other technical influences (**Fig. 1, 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 analyze the distribution of read depth for a genomic segment across many genomes using constrained Gaussian mixture models (**Fig. 1,2, Online Methods**). These models simultaneously infer the most-likely integer copy number for each genome and the confidence of each copy-number assignment, which we represent as copy-number “likelihoods” (the likelihood of each potential copy-number “genotype” given the sequencing data, analogous to genotype likelihoods for SNPs [28-30]). Population-level analysis also enables us to more accurately infer the correct range of absolute copy number 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 (versus looking at sites of known segmental duplication [22] or sites with other evidence of duplication, such as aberrantly oriented read pairs [31]). We scanned the human genome in overlapping windows, looking for segments where the read depth measurements deviated from a unimodal distribution (**Fig. 1, 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 individuals’ estimated copy numbers clustered at a series of integer levels (**Fig. 1, Online Methods**). For CNV loci that are 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 are unique genome-wide and positions that are identical between the duplicated segments on the reference genome (**Online Methods**). Such population-scale approaches become more powerful in the simultaneous analysis of many genomes (**Fig. 1,2**).

Using these approaches, we made a detailed ploidy map of 849 genomes from fourteen diverse human populations (**Supplementary Table 1**) from Phase 1 of the 1000 Genomes Project, sequenced at 4.8x median coverage (range 2.0x to 20.6x). 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 length10.7 kbp) of which 1,449 (16%) overlap protein coding genes. Some 4,781 of these CNVs appeared to be biallelic deletions; 2,522 were biallelic duplications; and 1,356 were multi-allelic (**Supplementary Note**). The distribution of observed diploid copy number at these loci ranged from zero to fifteen (**Fig. 2, 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 analyzes distributions of array probe intensity measurements across samples to estimate a false-discovery rate for the identification of CNV loci (**Online Methods**). This IRS test previously estimated false discovery rates (FDRs) of 2–83% for various deletion-discovery algorithms used in the 1000 Genomes Project[32]. For the CNV discovery set in the current work, this same test estimated a 2.7% FDR, including 5.8% for duplication CNVs overall and 1.7% for duplication CNVs larger than 10kb (**Supplementary Table 3**).

We next evaluated the specific integer copy-number determinations in each genome (often called “genotypes” in analogy to crisp SNP genotyping [7,15,33]). Importantly, 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 to the high-quality array-based analysis by Conrad *et al.* (**Supplementary Figure 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 Tables 4-5, Supplementary Figure 3, Supplementary Note). This is the first sequencing-based duplication call set to show such high, systematic 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 ranges from 0 to 5. To evaluate our genotypes for higher-copy-number CNVs, 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 [34,35]. We selected 22 high-copy CNVs with a wide dynamic range of copy number ranging from 1 to 9 (median 4, mean 4.47) and typed them using ddPCR in 90 HapMap samples. Integer genotype concordance was 99.9% (**Supplementary Table 6**) and consistent across the range of copy number evaluated (**Fig. 3, Supplementary Figure 4**). These data represent the first such resource for high-copy mCNVs (**Supplementary Data**).

## Alleles, phasing and haplotypes

Using these integer copy number determinations in 849 individuals, we next sought to infer copy number at the level of specific alleles or chromosomal copies. At mCNVs, the diploid copy 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 7 and 0.) Additional information exists at a population level, however, since the distribution of copy numbers contains information about the relative frequencies of different copy-number alleles [36,37]. There is also potential information in flanking SNP haplotypes that can potentially inform the analysis of individuals in whom multiple allelic combinations are possible[36,37].

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 beagle4 imputation software [38] 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 (**Supplementary Data, Supplementary Figure 5**).

This analysis revealed 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 three or more segregating alleles; 121 of these had four or more alleles, and 45 had five or more. (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 outside 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 are multi-allelic (35%) was greater than estimates from earlier studies. This is due to the larger sample size with its increased ability to ascertain low-frequency alleles; many duplication CNVs that we and others[7,15] 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 human genome's largest source of gene dosage variation

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 differs in 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 – contribute 88% of variation in human gene dosage (66 of the 75 gene copy-number differences present on average between any pair of individuals), approximately seven times more than the contribution of the more-numerous biallelic CNVs (**Table 1**). Several factors accounted for mCNV's large contribution to human gene dosage variation. First, duplications are more likely than deletions to affect protein-coding genes (**Table 1**), as reported previously[7]. Second, mCNVs contribute much more to gene-dosage variation than the more-numerous duplication CNVs do (**Table 1**). This latter result is largely due to allele frequency; duplication CNVs appear to have a propensity to become multi-allelic as they reach higher allele frequencies in populations (Table 1, Supplementary Figure 6, Supplementary Tables 7-8). We hypothesize that this is due to 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 gene dosage variation, 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 [22] and are likely to be multi-allelic.

Biallelic and multi-allelic 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 under-represented among genes involved in intracellular metabolic and biosynthetic pathways (**Supplementary Table 9, Supplementary Figure 7**).

## Impact on gene expression

Given the large contribution of mCNVs to human gene dosage variation, we sought to understand their contribution to gene expression variation. In the simplest model, increased dosage of a gene could directly cause increases in 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 [39-42]; or gene duplications might not include distal regulatory elements important for a gene's expression.

To address this question in one cell type, we utilized available mRNA-seq data derived from lymphoblastoid cell lines (LCLs) from 310 individuals whose genomes we had analyzed here [**43**] and asked whether variation in gene dosage is reflected in mRNA abundance. Most mCNV genes showed a strongly positive correlation between the gene's dosage in genomic DNA and that gene's mRNA expression (**Fig. 4** and **Supplementary Figure 8)**. RNA abundance tended to rise linearly with gene dosage (**Fig. 4**). Among 133 genes with evaluable expression levels in LCLs (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 RNA expression levels of the genes they contain. We further explore relationships between mCNVs and gene expression in **Supplementary Table 10** and **Supplementary Figure 9**.

## 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 six common diseases (1,500 cases per disease) by the WTCCC found few effects from common

CNVs [44]; however, most discoveries of SNP association for complex phenotypes have required much larger samples (tens of thousands) and been made during meta-analysis of many GWAS data sets. Such discovery 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 [35] and is somewhat tractable to imputation [35]. 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 beagle4[38] software, we evaluated how effectively diploid copy number can 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 results 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 MAF > 10%) exhibiting almost an almost uniform distribution of imputation $r^2$ from 0 to 1.0 (**Supplementary Figure 10**). This suggests that mCNVs span a wide range, from loci at which discrete copy-number alleles are well-imputed, to loci with recurring mutations and little sustained relationship of copy number to surrounding haplotypes.

A scenario that could potentially hinder the imputation of some mCNVs would occur if the duplication alleles were dispersed to distant genomic sites[45,46]. We looked for evidence of dispersed duplications using long-range LD and interchromosomal segmental duplications and found 15 CNVs with evidence of dispersal (Supplementary Tables 11-12, Supplementary Figure 11); 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” (**Fig. 5d**) (i.e. when a mCNV's less-common alleles have relatively high frequency). 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 these genes was *HPR*, which encodes a haptoglobin-related protein that is used in defense against trypanosomes[47]. *HPR* was present at 2 copies in European population samples and 1-2 copies in Asian population samples, but in 4-8 copies in about 25% of individuals sampled from African populations (**Fig. 6**). A second example was the *ORM1* (orosomucoid) gene, 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 low- and high-copy number alleles were segregating. At both *ORM1* and *HPR*, the high-copy-number alleles were observed on a shared SNP haplotype, while 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-number alleles suggest a recent, geographically unique origin. These relationships also indicate that these high-copy alleles have evolved by repeated, recurrent mutation on the same haplotype background.

Both *HPR* and *ORM1* are functionally connected to disease loci. The *HPR* protein forms (together with *APOL1*) the Trypanosome Lytic Factor (TLF) that is a primary defense of human blood against trypanosomes[47,48]. *APOL1* appears to have been under strong recent selection in African populations due to recently-arisen variants that protect against trypanosome infection while contributing to kidney disease[49]. The copy-number variation at *HPR* is similar to the Africa-specific SNPs at *APOL1* in that it contains multiple new alleles that have quickly risen to high frequencies in African populations where trypanosomes are endemic[49,50]. *ORM1*, one of the most abundant glycoproteins in blood, is paralogous to *ORMDL3*, in which common variants associate strongly to risk of asthma[51]. It is intriguing to speculate that *HPR* and/or *ORM1* may have quickly risen 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 distinguish the high-copy-number alleles from the more-common low-copy-number alleles (**Fig. 6**). However, SNP haplotypes do not readily distinguish the high-copy alleles from one another (**Fig. 6**). It seems likely that high-copy-number alleles have frequently mutated into other high-copy-number states, while low-copy-number alleles have remained stable. Understanding the actual mutation rates of mCNVs is an important direction for future work.

## Discussion

We described computational approaches for identifying mCNVs and characterizing such polymorphisms in terms of alleles, allele frequencies, and haplotypes. The resulting data showed that mCNVs broadly impact genes and gene expression. Multi-allelic CNVs give rise to at least six times more gene-dosage variation than simpler, biallelic CNVs do (**Table 1**), and such variation usually affects the expression level of the encompassed genes (**Fig. 4**).

Multi-allelic CNVs are not routinely evaluated in genome-wide studies based on SNP arrays or exome sequencing due to the limitations of these approaches for accurately measuring copy numbers greater than four. Two studies have analyzed mCNVs using custom-designed targeted arrays, which appear to be more effective, though 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[44,52]. Several features of mCNVs suggest that mCNVs 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 to mCNVs will exhibit a clear direction of effect (i.e. certainty about whether more or less gene activity contributes to risk). Third, therapeutics based on such discoveries may 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? Increasingly abundant WGS data, together with analysis methods such as those here, could increasingly make direct association testing routine. But it will take many years (and substantial resources) for the sample size of WGS-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. Based on observing at least nominal association to 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 to mCNV copy number at many loci could empower conditional association analysis to disentangle the effects of mCNV alleles from effects 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 here at imputing 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 to date required customized computational and molecular strategies[35,45,53,54].

The elucidation of complex and recurrently mutating forms of genome variation will ultimately deepen our 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 (including plots of flanking haplotypes) are available at http://www.broadinstitute.org/software/genomestrip/mcnv_supplementary_data.

# Online Methods

## CNV discovery and genotyping

Discovery and genotyping of CNVs was performed using an enhanced version of the 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 base-pair reads can be uniquely aligned) and one targeting segmentally duplicated regions where the copy number in the reference genome is greater than one. For the uniquely-alignable portion of the genome, we utilized a pipeline which prospectively genotyped overlapping windows across the genome, using a window size of 5kb of alignable sequence and overlapping adjacent windows by 2.5kb (Discovery Set 1, **Supplementary Note**). For segmentally duplicated regions, we utilized 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 two, then subsequently filtering 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 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 is the midpoint of the left-most read (for paired-end sequencing). Reads were only counted if they mapped to locations which should be uniquely alignable based on 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 browser). The raw read counts were normalized to correct for sequencing bias as a function of the G+C content in 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 Gaussians, corresponding to a series of diploid copy number classes from zero to a site-specific maximum. The maximum copy number modeled for each site was chosen based on 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 were 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 at copy number 4-6. To avoid over-fitting, we rejected models where the scaling parameter is too high (above 2.0) or too low (below 0.5).

## Utilizing copy number parity

To further increase the accuracy of absolute copy-number determination, especially at high copy number, we borrowed information from the population allele frequencies of the samples being jointly analyzed. At autosomal loci, under assumption of Hardy-Weinberg equilibrium (HWE), the number of individuals that 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 in 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 HWE. 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 exploit this information by performing a simple parity test on copy number. As we fit the mixture model, we estimate the number of even and odd copy numbers observed in the population. If the fraction of even copy numbers falls below a specified threshold parameter (default 0.4), we shift the cluster assignments towards 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 that are sufficiently unique that sequence reads should be able to align uniquely (based on 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 utilize positions that are non-unique on the reference 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 utilized together to estimate both the total copy number genome-wide and the paralog-specific copy number. The unique positions represent paralog-specific differences (or paralog-specific variation, PSVs) that differentiate one paralog from the other, based on 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 converges most strongly to 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 (200bp) or minimum segment length (2.5kb) was reached.

### CNV phasing and imputation

Prior to 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 (e.g. 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 and considering 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 (beagle4, version r1128) and the 1000 Genomes Phase 1 reference panel supplied with the beagle software. Each CNV site was phased and imputed separately. For CNVs at segmentally duplicated sites where the segments were separated by more than 100kb, 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 10 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 10 individuals at a time. The imputed genotypes were compared to the measured diploid copy number for the withheld individuals, using 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 (intensity rank sum) test implemented in the Genome STRiP software package to evaluate false discovery rate. This test ranks the normalized SNP array probe intensity values from the SNPs underneath 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 false discovery rate was estimated from the distribution of p-values across the evaluated sites, considering deletions and duplication sites separately (**Supplementary Note**). The overall call set FDR 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 (Yoruba in Ibidan, Nigeria) cohort of the Coriell 1000 Genomes Project samples for 32 multi-allelic CNV and segmental duplication target sites ascertained from sequencing data (**Supplementary Table 13**) using droplet digital PCR (ddPCR, Bio-Rad Laboratories) and standard protocols (**Supplementary Note**). Primer/probe ddPCR assays for each target were custom designed by Bio-Rad Laboratories based on 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 was downloaded from the Geuvadis project web site. For CNVs that fully overlapped genes, we computed the Pearson's correlation coefficient between the CNV integer copy number and the normalized gene expression quantitation from the RNA sequencing data.

P-values for gene expression were calculated using 10,000 random permutations and a one-sided test with 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.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported by a grant from the National Human Genome Research Institute (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. We thank Dimitri Skvortsov, Michael

Thornton, Niels Klitgord, and Bin Zhang for contributions to ddPCR assay design and validation. We would also like to thank members of the 1000 Genomes Project for helpful conversations about analysis methods.

## References

1. Sebat J, et al. Strong association of de novo copy number mutations with autism. Science. 2007; 316:445–9. [PubMed: 17363630]
2. International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 2008; 455:237–41. [PubMed: 18668038]
3. Weiss LA, et al. Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med. 2008; 358:667–75. [PubMed: 18184952]
4. McCarthy SE, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009; 41:1223–7. [PubMed: 19855392]
5. Bochukova EG, et al. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature. 2010; 463:666–70. [PubMed: 19966786]
6. Vacic V, et al. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature. 2011; 471:499–503. [PubMed: 21346763]
7. Conrad DF, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–12. [PubMed: 19812545]
8. McCarroll SA, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genetics. 2008; 40:1166–1174. [PubMed: 18776908]
9. de Cid R, et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet. 2009; 41:211–5. [PubMed: 19169253]
10. McCarroll SA, et al. Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. Nat Genet. 2009; 41:1341–4. [PubMed: 19935662]
11. Willer CJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet. 2009; 41:25–34. [PubMed: 19079261]
12. McCarroll SA, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet. 2008; 40:1107–12. [PubMed: 19165925]
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. 2007; 17:1665–74. [PubMed: 17921354]
14. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature Genetics. 2011; 43:269–U126. [PubMed: 21317889]
15. McCarroll SA, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008; 40:1166–74. [PubMed: 18776908]
16. Hollox EJ, Armour JA, Barber JC. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet. 2003; 73:591–600. [PubMed: 12916016]
17. Iafrate AJ, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36:949–51. [PubMed: 15286789]
18. Lee C, Iafrate AJ, Brothman AR. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet. 2007; 39:S48–54. [PubMed: 17597782]
19. Perry GH, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007; 39:1256–60. [PubMed: 17828263]
20. Perry GH, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. 2008; 82:685–95. [PubMed: 18304495]
21. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008; 1:4. [PubMed: 19014668]
22. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. Science. 2010; 330:641–6. [PubMed: 21030649]
23. Alkan C, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009; 41:1061–7. [PubMed: 19718026]

24. 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. 2009; 19:1586–92. [PubMed: 19657104]

25. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21:974–84. [PubMed: 21324876]

26. Bellos E, Johnson MR, LJ MC. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. Genome Biol. 2012; 13:R120. [PubMed: 23259578]

27. Koren A, et al. Genetic variation in human DNA replication timing. Cell. 2014

28. Wang Y, Lu J, Yu J, Gibbs RA, Yu F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 2013; 23:833–42. [PubMed: 23296920]

29. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–8. [PubMed: 21478889]

30. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011; 27:2987–93. [PubMed: 21903627]

31. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28:I333–I339. [PubMed: 22962449]

32. Mills RE, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011; 470:59–65. [PubMed: 21293372]

33. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007; 39:S37–42. [PubMed: 17597780]

34. Hindson BJ, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem. 2011; 83:8604–10. [PubMed: 22035192]

35. Boettger, LM.; Handsaker, RE.; Zody, MC.; McCarroll, SA. Nature Genetics. Vol. 44. 881-+: 2012. Structural haplotypes and recent evolution of the human 17q21.31 region..

36. Su SY, et al. Inferring combined CNV/SNP haplotypes from genotype data. Bioinformatics. 2010; 26:1437–45. [PubMed: 20406911]

37. Kato M, Nakamura Y, Tsunoda T. An algorithm for inferring complex haplotypes in a region of copy-number variation. Am J Hum Genet. 2008; 83:157–69. [PubMed: 18639202]

38. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009; 84:210–23. [PubMed: 19200528]

39. Assaad FF, Tucker KL, Signer ER. Epigenetic repeat-induced gene silencing (RIGS) in Arabidopsis. Plant Mol Biol. 1993; 22:1067–85. [PubMed: 8400126]

40. Dorer DR, Henikoff S. Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell. 1994; 77:993–1002. [PubMed: 8020105]

41. Dorer DR, Henikoff S. Transgene repeat arrays interact with distant heterochromatin and cause silencing in cis and trans. Genetics. 1997; 147:1181–90. [PubMed: 9383061]

42. Garrick D, Fiering S, Martin DI, Whitelaw E. Repeat-induced gene silencing in mammals. Nat Genet. 1998; 18:56–9. [PubMed: 9425901]

43. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–11. [PubMed: 24037378]

44. Wellcome Trust Case Control, C. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010; 464:713–20. [PubMed: 20360734]

45. Abu Bakar S, Hollox EJ, Armour JA. Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. Proc Natl Acad Sci U S A. 2009; 106:853–8. [PubMed: 19131514]

46. Dennis MY, et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell. 2012; 149:912–22. [PubMed: 22559943]

47. Smith AB, Esko JD, Hajduk SL. Killing of trypanosomes by the human haptoglobin-related protein. Science. 1995; 268:284–6. [PubMed: 7716520]

48. Harrington JM, Howell S, Hajduk SL. Membrane permeabilization by trypanosome lytic factor, a cytolytic human high density lipoprotein. J Biol Chem. 2009; 284:13505–12. [PubMed: 19324878]

49. Genovese G, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010; 329:841–5. [PubMed: 20647424]

50. Genovese G, Friedman DJ, Pollak MR. APOL1 variants and kidney disease in people of recent African ancestry. Nature Reviews Nephrology. 2013; 9:240–244.

51. Moffatt MF, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448:470–3. [PubMed: 17611496]

52. Zanda M, et al. A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes. PLoS Genet. 2014; 10:e1004367. [PubMed: 24875393]

53. Hollox EJ, et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet. 2008; 40:23–5. [PubMed: 18059266]

54. Steinberg KM, et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet. 2012; 44:872–80. [PubMed: 22751100]

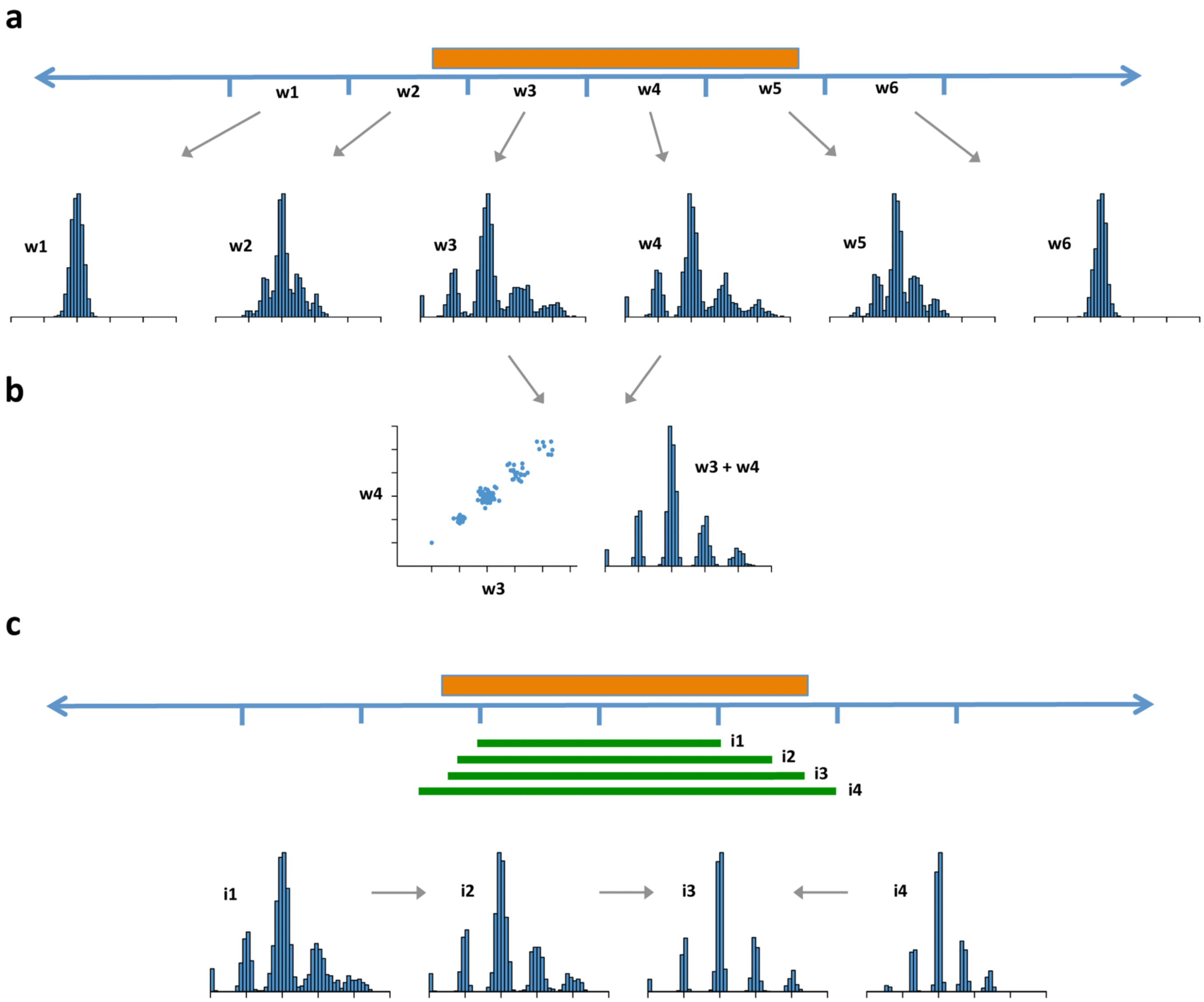

**Figure 1.**

Ascertainment of multi-allelic copy number variations (mCNVs) across the human genome. Multi-modal patterns of variation for a high-frequency CNV (orange box represents the true extent of the CNV) can be detected in multiple windows (w1 – w6) that overlap the CNV segment **(a)**. Where read-depth distributions from adjacent windows are highly correlated across many genomes, these windows are merged to increase power for genotyping **(b)**. To more precisely estimate the genome sequence affected **(c)** 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).

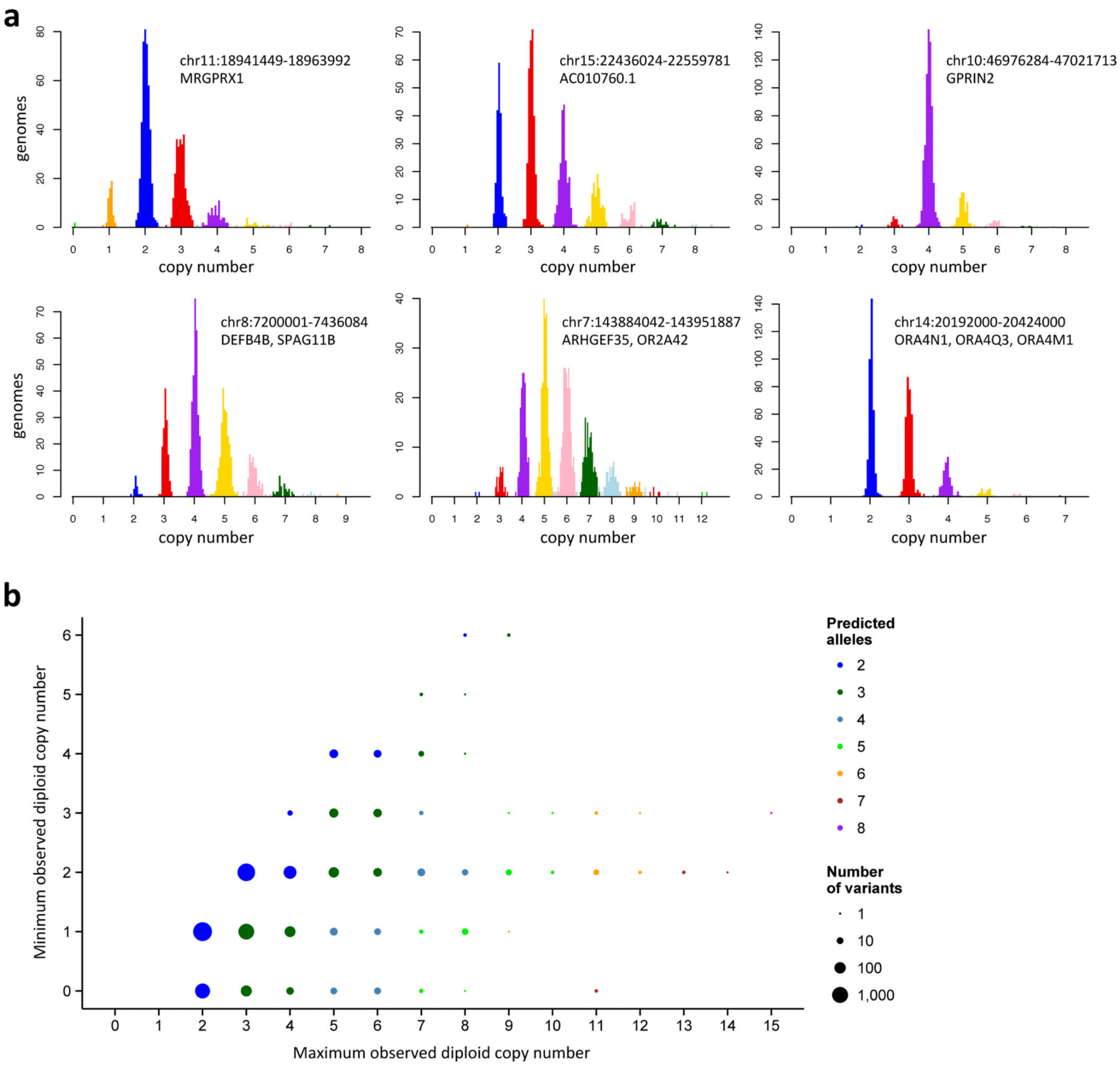

**Figure 2.**

Determination of the copy-number levels and alleles present at mCNV loci. Histograms of normalized read depth **(a)** are fitted with a Gaussian mixture model to infer 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. Similar plots for all mCNVs ascertained in this study are provided in the supplementary web resource. **(b)** Distribution of observed diploid copy numbers across 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 represent the observed dynamic range of copy number variation observed at each site. For example, the blue circles represent deletions and duplications, and the various other colors represent classes of multi-allelic CNVs with various copy-number ranges and numbers of alleles.

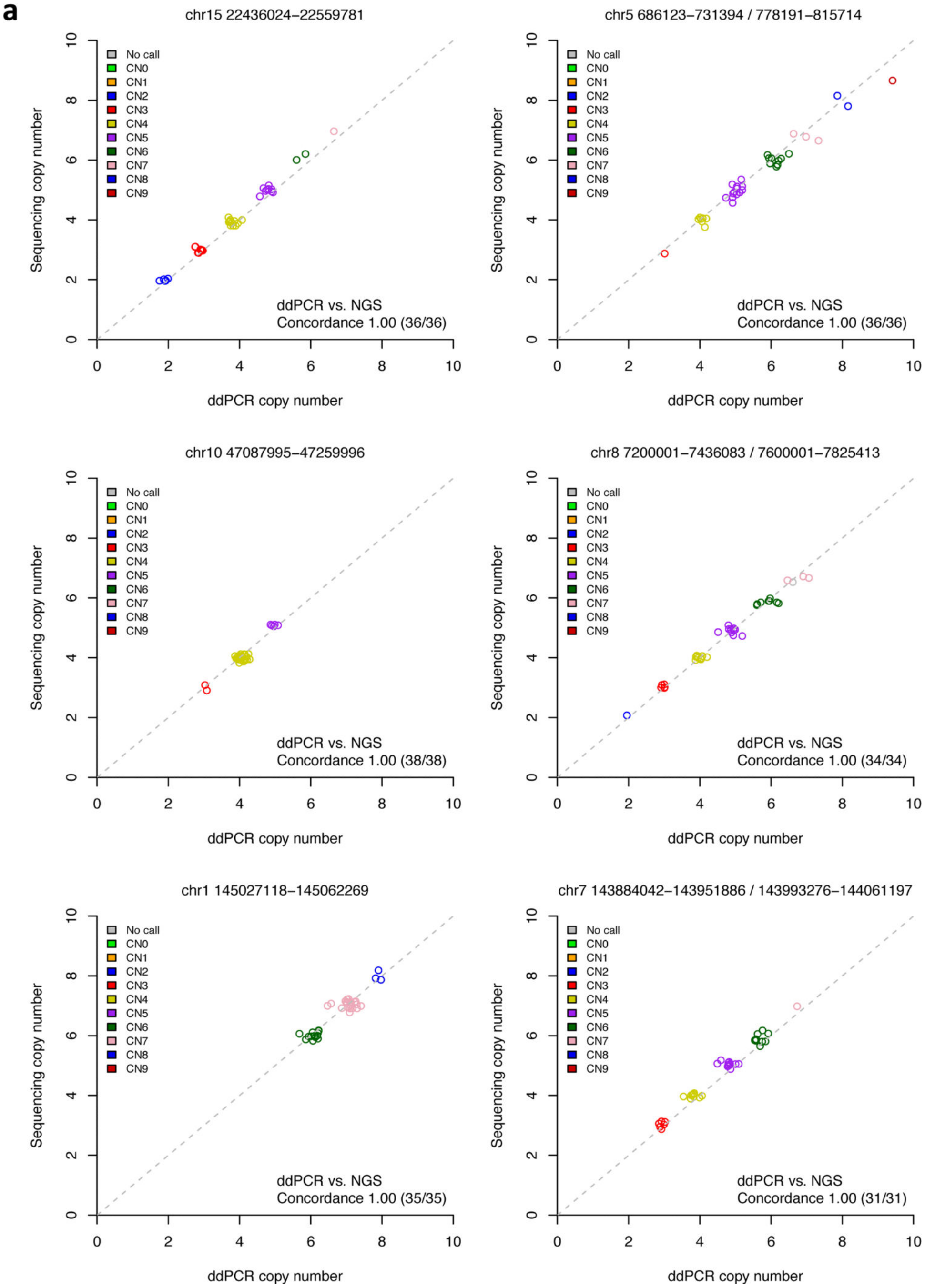

**Figure 3.**

Critical evaluation of copy number genotypes by droplet digital PCR (ddPCR). Across 38 genomes evaluated, copy number genotypes from sequencing data were compared with measurements from ddPCR. Panel **(a)** shows data for 6 of the 22 loci evaluated. Plots for all loci are in Supplementary Figure 4. Across the 22 loci, the two methods showed 99.9% genotype concordance at confidently called sites.

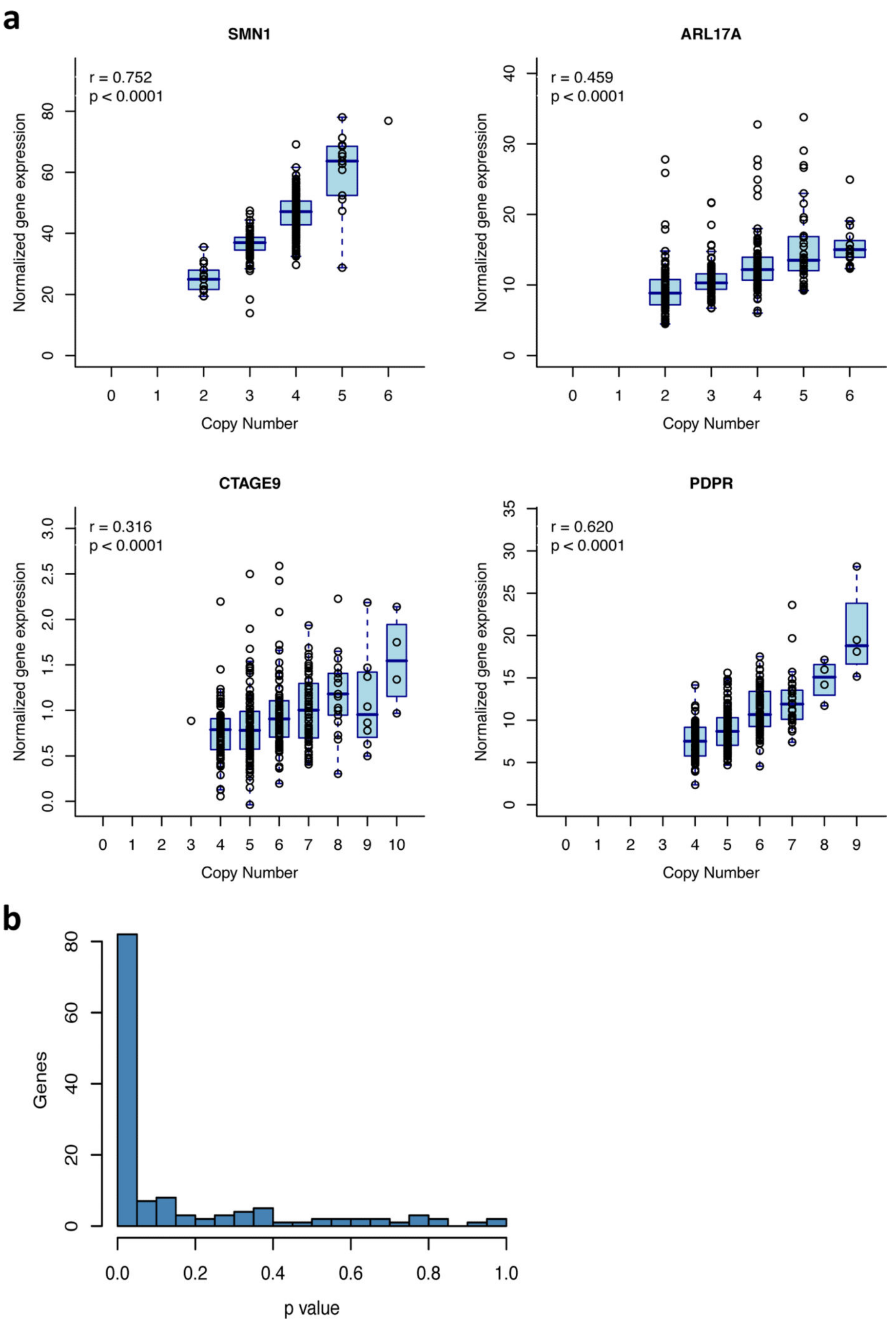

**Figure 4.**

Relationship of gene copy number (in genomic DNA) to gene expression (in mRNA) for multi-allelic CNVs. **(a)** At four typical genic mCNVs, inter-individual variation in gene expression levels appears to arise strongly from gene dosage variation. For clarity, individual points show data for all 310 genomes that have corresponding RNA data (not just outliers) overlaid on a summary box plot showing median, inter-quartile range, and whiskers extending to the most extreme point no more than 1.5 IQR from the box edge (Tukey convention). **Supplementary Figure 8** shows such analyses for many more genes. **(b)**

Across all genic mCNV loci that are expressed in lymphoblastoid cell lines, the distribution of p-values (in tests for positive correlation between gene expression and gene dosage) is dominated by low p-values. P-values calculated using 10,000 random permutations (**Supplementary Note**).

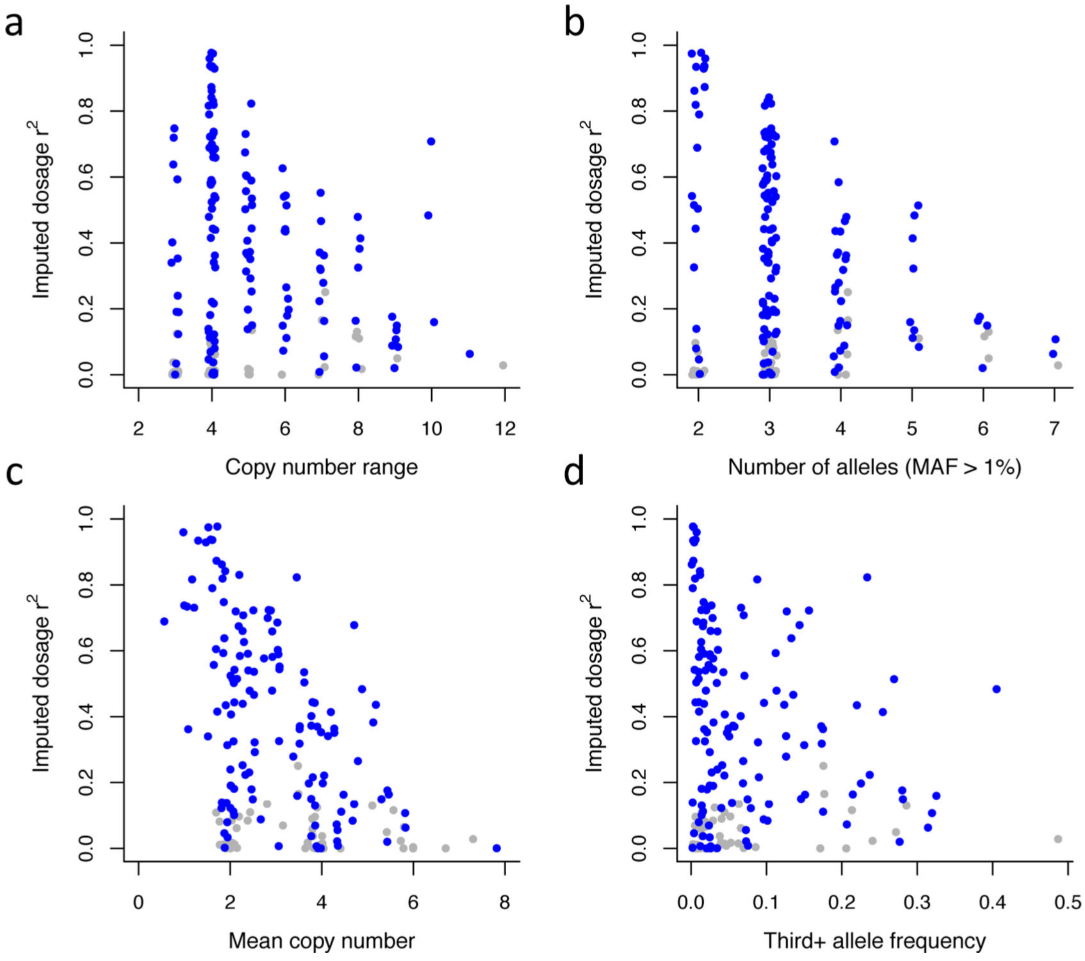

**Figure 5.**

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

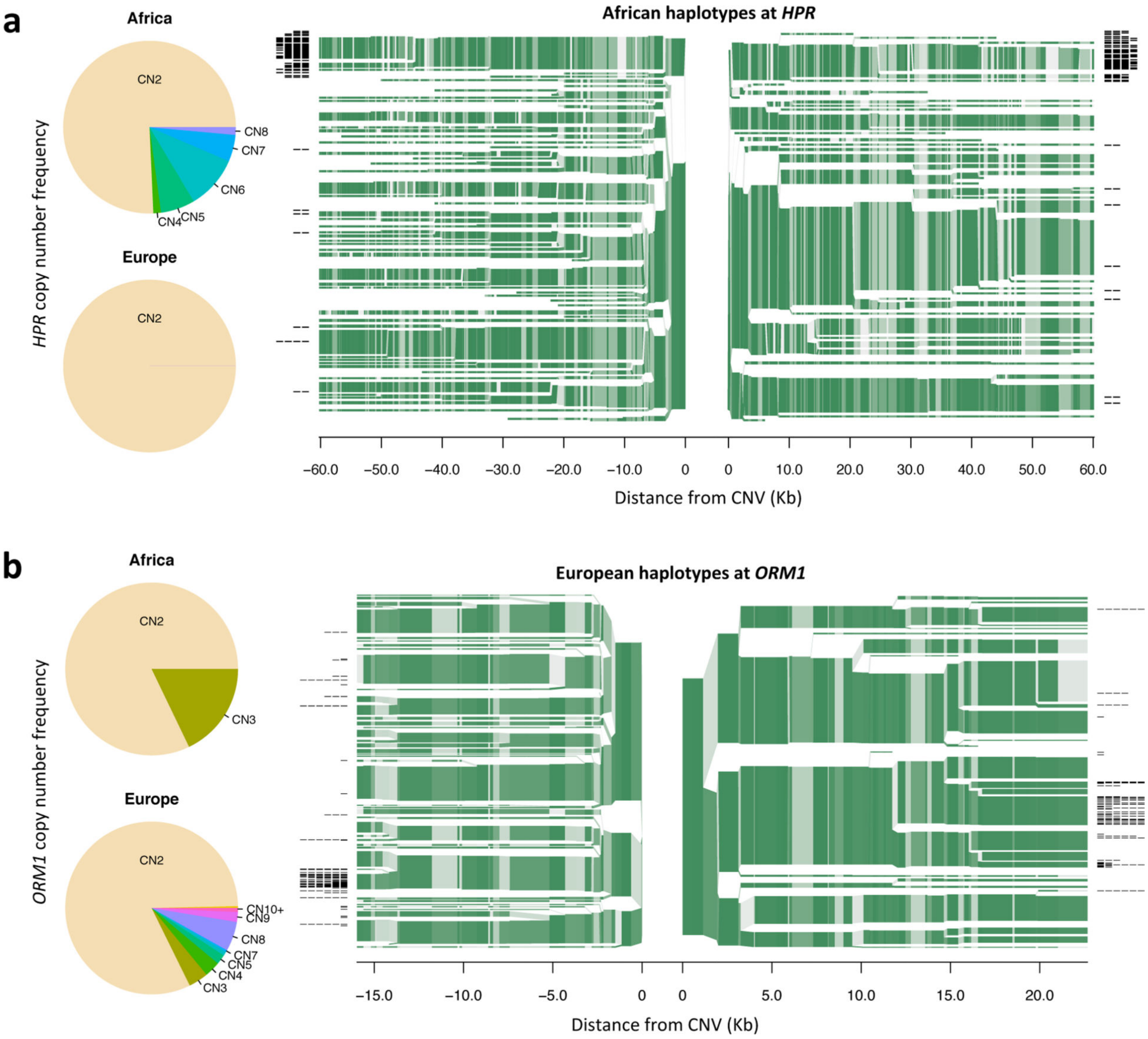

**Figure 6.**

Haplotypes with “runaway” copy number. **(a)** Copy number distribution and haplotype structure of a multi-allelic CNV 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, in chromosomes sampled from African populations (YRI and LWK). 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 due to mutation or recombination. The thickness of each branch indicates haplotype frequency; shading indicates allele frequency of the individual SNPs used to define haplotypes. Haplotypes carrying high-copy *HPR* alleles (with more than one *HPR* copy) are indicated by black lines at branch tips with a line segment for each extra copy above one. Almost all the high-copy alleles appear to segregate on the same haplotype background. **(b)** A similar analysis of a mCNV affecting the *ORM1* gene, which appears to have greatly expanded in copy number on a specific haplotype, producing many different high-copy alleles.

## Table 1

CNV impact on gene dosage

| | | | | Number of genes differing in copy number between two individuals | | | | |
|---|---|---|---|---|---|---|---|---|
| **Category** | **# CNVs** | **Genic CNVs** | **Genes** | **Singletons** | **AAF < 1%** | **AAF < 5%** | **AAF > 5%** | **Overall** |
| Deletions | 4,781 | 70 | 88 | 0.12 | 0.33 | 0.54 | 5.04 | 5.58 |
| Duplications | 2,522 | 194 | 314 | 0.40 | 1.28 | 2.32 | 1.06 | 3.38 |
| Multi-allelic | 1,356 | 126 | 231 | n/a | 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 multi-allelic CNVs) to gene dosage variation among humans. Multi-allelic CNVs, which comprise only about 15% of the CNVs ascertained, give rise to more than 85% of human gene dosage variation, 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.