ARTICLE

Identification of Copy Number Variation Hotspots in Human Populations

Wenqing Fu,[1] Feng Zhang,[1] Yi Wang,[1] Xun Gu,[1,2] and Li Jin[1,3,*]

Copy number variants (CNVs) in the human genome contribute to both Mendelian and complex traits as well as to genomic plasticity in evolution. The investigation of mutational rates of CNVs is critical to understanding genomic instability and the etiology of the copy number variation (CNV)-related traits. However, the evaluation of the CNV mutation rate at the genome level poses an insurmountable practical challenge that requires large samples and accurate typing. In this study, we show that an approximate estimation of the CNV mutation rate could be achieved by using the phylogeny information of flanking SNPs. This allows a genome-wide comparison of mutation rates between CNVs with the use of vast, readily available data of SNP genotyping. A total of 4187 CNV regions (CNVRs) previously identified in HapMap populations were investigated in this study. We showed that the mutation rates for the majority of these CNVRs are at the order of 10^{-5} per generation, consistent with experimental observations at individual loci. Notably, the mutation rates of 104 (2.5%) CNVRs were estimated at the order of 10^{-3} per generation; therefore, they were identified as potential hotspots. Additional analyses revealed that genome architecture at CNV loci has a potential role in inciting mutational hotspots in the human genome. Interestingly, 49 (47%) CNV hotspots include human genes, some of which are known to be functional CNV loci (e.g., CNVs of *C4* and β-defensin causing autoimmune diseases and CNVs of *HYDIN* with implication in control of cerebral cortex size), implicating the important role of CNV in human health and evolution, especially in common and complex diseases.

Introduction

Recent studies have shown that the presence of copy number variants (CNVs) in the human genome is substantial, and thousands of CNVs have been identified in human populations.[1,2] CNV mutations can introduce unprecedented genomic instability in both germline and somatic cells, which would lead to Mendelian diseases and complex traits, including cancers.[3,4] Investigation of the CNV mutation rate is, therefore, critical to unraveling the instability of the human genome and in turn the functional impact of CNVs underlying human traits and diseases.

The inbreeding across hundreds of generations made it accessible to study the mutation rate for spontaneous CNVs in the laboratory mouse strains.[5] However, the direct estimation of germline mutation rate for CNVs in the human genome is technically challenging. Only a few copy number variation (CNV) loci have been studied individually by disease-prevalence calculation, pedigree analysis, sperm typing, or pooled sperm assay,[6–11] but a systematic investigation of the CNV mutation rate across the genome has yet to be accomplished. The locus-specific mutation rates for CNV have been observed to be ~100 to 10,000 times higher than those for nucleotide substitution rates,[3] which not only highlights the instability of CNV regions but also suggests large variation in CNV mutation rate.

In order to examine whether such a notable variance of mutation rate observed between CNV loci is potentially due to the inconsistency among various locus-specific investigatory mothods[6–8,10,11] or instead reflects the presence of CNV mutational hotspots, a genome-wide analysis for estimation and comparison of CNV mutation rate is demanded. In this study, we developed a statistical method called "CNVMut" to evaluate CNV mutation rate across autosomes of the human genome by utilizing vast, readily available genotyping data from human populations. This proposed method was carefully evaluated and validated via simulation approaches, and it was applied to identifying mutational hotspots of CNVs in HapMap populations.

Material and Methods

An Algorithm for Approximately Estimating the CNV Mutation Rate by Using Flanking SNPs

CNVs can be classified as recurrent CNVs with common breakpoints and as nonrecurrent CNVs with variable breakpoints.[3] A CNV region (CNVR) is usually defined to combine calls from different individuals as a grouping of CNVs overlapping or in close proximity to each other, regardless of the potential architectural complexity.[2,12] In this study, a method was proposed for approximate estimation of the mutation rate for each CNVR. Each CNVR was treated as a simple marker with potentially multiple alleles, each with a distinct copy number (0, 1, 2, etc.). Each individual carries two such alleles in his or her diploid genome.

The SNPs flanking CNV loci often show strong linkage disequilibrium (LD) with the CNVs,[1,12–14] therefore allowing one to trace CNV mutation events by using the phylogenies of flanking SNPs constructed by plausible ancestral recombination graphs

[1]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China; [2]Department of Genetics, Development and Cell Biology-Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA; [3]Department of Computational Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
*Correspondence: ljin007@gmail.com
DOI 10.1016/j.ajhg.2010.09.006.

(ARGs).[15] ARGs can describe the relationship of chromosome sequences from a population back to its common ancestor through events of mutation, recombination, and coalescence, which are defined under a Wright-Fisher model with recombination. According to the impact of historical recombination events, a series of genealogical trees are embedded in the ARG. Each genealogical tree describes the phylogeny for each SNP, which is also called a "marginal tree." Following each marginal tree for the flanking SNPs, we define the tree with the smallest number of CNV changes as a surrogate of the most parsimonious genealogical tree for the CNV locus according to the parsimony criterion. The number of CNV mutations (including both deletion and duplication) is summarized on the basis of the most parsimonious tree and can be converted to an approximate estimation of mutation rate for the CNV locus by simulations.

A brief summary of our statistical method for indirectly estimating the CNV mutation rate is illustrated in Figure S1 (available online). First, the program Margarita[15] was implemented for the construction of plausible ARGs by the SNPs flanking each CNV locus in one population. The corresponding marginal tree for each SNP was obtained in this step as well, the external nodes representing the inferred haplotypes of genotyped individuals, the internal nodes representing the ancestral haplotypes in previous generations, and the root representing the common ancestral haplotype. Second, an expectation-maximization (EM) algorithm was used for inferring haplotypes involving a CNV marker and its flanking SNPs as described by Kato et al.,[16] and we also modified the algorithm to allow the input of phase-known data (i.e., the preinferred haplotypes of flanking SNPs obtained from ARG construction). Third, haploid copy numbers of a CNV marker were superimposed to corresponding external nodes that represented the preinferred haplotypes of flanking SNPs for each marginal tree, and they were inferred for each internal node by traversing all the binary tree units (BTUs; Figure S1D). Lastly, the number of mutation events was counted for all BTUs, the smallest count being taken as the most parsimonious estimator. In particular, deletion events and duplication events were counted separately along the most parsimonious marginal tree, with the assumption that a tree requiring fewer substitutions is better than one that requires more, for evaluation of the relative mutation rate between deletion and duplication. This detailed procedure is described in Appendix A.

Because the real phylogeny was unknown, the average of the estimator for R plausible ARGs ($R = 100$ here) was taken. For example, the cumulative mutation count, M, was estimated by taking an average of the inferred number of mutations over R ($R = 100$ here) plausible ARGs:

$$M = \frac{\sum_{r=1}^{R} \min_{l=1}^{L}\{M_{rl}\}}{N * R},$$

where M_{rl} is the estimator of the mutation number for the l-th marginal tree of the r-th plausible ARG, and N is the sample size.

The statistic M measures the minimum effective number of mutation events instead of mutation rate in a population, following the parsimony assumption; i.e., the estimation depends on the observed cumulative mutation events under an assumption of constant evolution rate. It allows an indirect comparison of mutation rates between CNV loci and between populations. The order of magnitude of the mutation rate could be converted from the M statistic with the use of a simulation approach.

Simulation in Evaluating the Algorithm Performance

Given that the aforementioned algorithm invoked several assumptions that may not biologically realistic, it was evaluated via a simulation approach. In particular, we interrogated the effect of several factors on the estimation of M, including mutation rate, effective population size, number of flanking SNPs, sample size, recombination rate between the CNV locus and its flanking SNPs, and that between adjacent SNPs. The program SIMCOAL2[17] was employed in simulating the neutral genetic diversity of a CNV marker and its flanking SNPs under different mutation rates and demographic models (see Appendix B).

Algorithm Implementation for Real Data

Genotype data for the unrelated samples from the HapMap Phase II project were included in our analysis: 60 Yoruba in Ibadan, Nigeria (YRI), 60 Utah residents with ancestry from northern and western Europe (CEU), and 89 Asians (45 Han Chinese from Beijing [CHB] and 44 Japanese from Tokyo [JPT]; CHB+JPT). First, 4330 genotyped CNVRs from 22 autosomes, identified by a previous study in HapMap populations, were investigated.[2] Genotypes of the SNPs flanking CNV regions but not involved in CNVR were obtained from the HapMap website (NCBI build 36, release 24). Second, the genetic distance between each CNVR and its nearest upstream and downstream SNPs was calculated according to the local recombination rate estimation obtained from the HapMap website (NCBI build 36) (Figure S2). The CNVRs with a genetic distance of > 0.1cM to its flanking SNPs were excluded from further analysis, so as to minimize the effect of high recombination rate between a CNVR and its flanking SNPs on the estimation of M. Overall 4187 CNVRs were enrolled in the following analyses.

The proposed method, CNVMut, was implemented to estimate both the statistic M and the proportion of deletion mutation for each CNV locus by using a total of 20 flanking SNPs for each population, respectively. Pairwise comparisons of the estimates of M among three HapMap populations were conducted by Spearman's correlation test with the use of all of the CNVs and/or the exclusion of those CNVs with the largest decile of *Fst* defined by Weir and Cockerham.[18] *Fst* was calculated according to allele frequencies estimated during the CNV allele inference by EM algorithm.

For comparison of the mutation rate of CNVs between populations and detection of mutational hotspots of CNVs, coalescent simulations were employed in converting the estimates of M to different orders of the mutation rate. The simulated data with different mutation rates of CNVs at order of magnitudes (5×10^{-7}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, and 5×10^{-3} per generation) were respectively produced by an extension of SIMCOAL2 under the proper selection of a demographic model for each population.[19] The demographic model used in coalescent simulations for three populations (YRI, CEU, and CHB+JPT) is illustrated in Table S1. In each simulation under a proper demographic model, individuals (60 for the YRI model, 60 for the CEU model, and 89 for the CHB+JPT model) with a CNV at a given mutation rate and 20 flanking SNPs with a minor allele frequency (MAF) > 0.01 were simulated. The recombination rate of adjacent loci, including that between a CNV marker and its flanking SNPs and that between adjacent SNPs, was assumed to be 10^{-5} per generation, based upon the evidence including the following: (1) the estimation of M is robust when the recombination rate between adjacent loci is less than 10^{-3} per generation (Figures S3E and S3F); (2) the

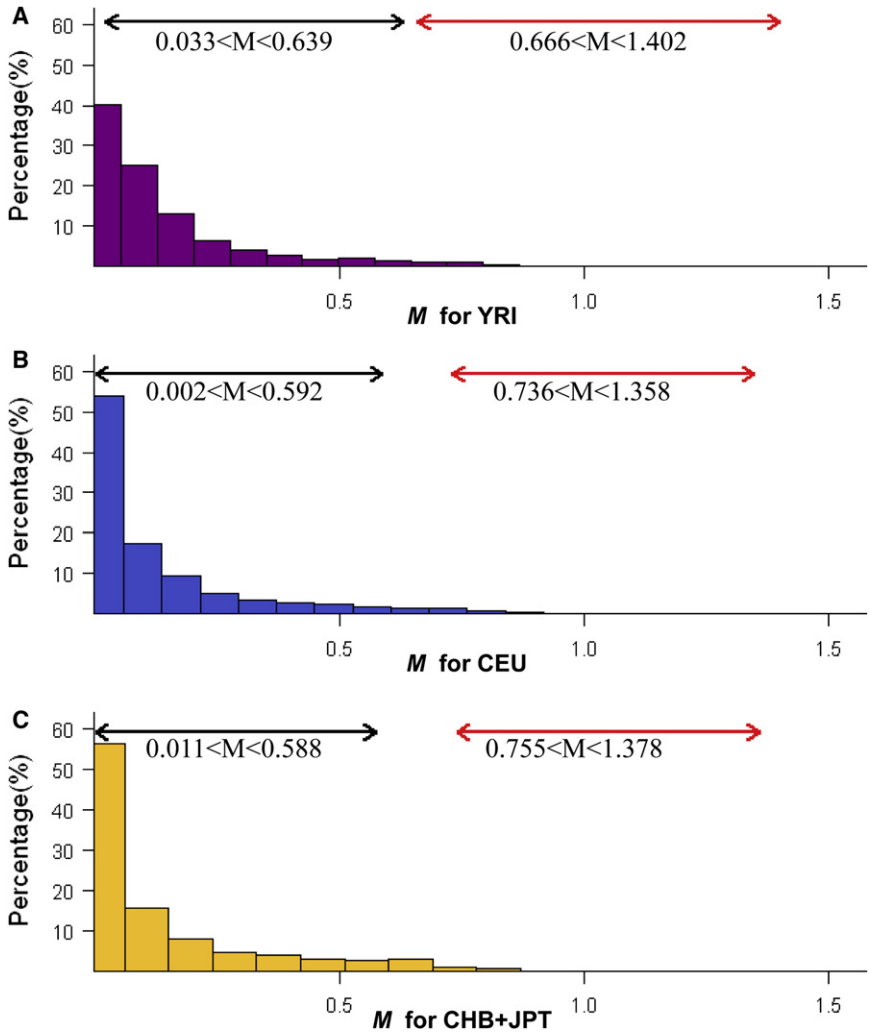

Figure 1. Distribution of the Estimates of *M* in Three HapMap Populations
The distribution of the estimates of *M* for 3147 polymorphic CNVRs in YRI (A), 2360 polymorphic CNVRs in CEU (B), and 1615 polymorphic CNVRs in CHB+JPT (C) is plotted. The range of *M* corresponding to a given mutation rate is indicated in the upper part of each plot (black for ~10^{-5} per generation and red for ~10^{-3} per generation). The range of *M* for a given mutation rate indicates the 2.5–97.5 percentile, based on 1000 coalescent simulations, which are based upon properly selected demographic models for each population.

genetic distances between CNV loci to their flanking SNPs were less than 0.1cM and 3.96 × 10^{-3} cM on average; and (3) the average recombination rate of the human genome is ~10^{-8} per generation per bp,[20,21] and the SNP density of HapMap Phase II is 1.14 per kb.[22]

Mutational hotspots for CNVs were identified according to the approximate estimation of the mutation rate converted from the estimates of *M*. Some potential differentiating features between mutational hotspots and the remaining loci, including the adjacency to segmental duplications (SDs; alternatively termed low-copy repeats [LCRs]),[23,24] the recombination rate between the CNVR and its flanking SNPs, the CNVR size and the mechanisms underlying CNV formation, were compared by Fisher's exact test or the Mann-Whitney test. The location information of SDs was obtained from the UCSC Genome Browser (hg 18) for examining whether CNVs overlapped with SDs or not. The recombination rate between the CNVR and its flanking SNPs was obtained from the HapMap website (NCBI build 36). The genetic diversity of the CNVR was measured by the inferred heterozygosity according to allele frequencies estimated by the EM algorithm. The CNV formation mechanisms were obtained from Conrad et al.[2]

In addition, the potential association of the relative contribution of deletion versus duplication in the mutation rate with the different CNV formation mechanisms was also investigated by the Mann-Whitney test.

Results

Evaluation of the Performance of the Estimation of *M* Using Simulated Data

The performance of the statistic *M*, under both circumstances of considering demographic events or not, was carefully evaluated and validated via simulation approaches. Simulations illustrated that the estimates of *M* increase with increasing mutation rates and can be affected by the demographic model. However, the number of flanking SNPs used in constructing ARGs, sample size, and recombination rate between adjacent loci (especially when recombination rate is $< 10^{-3}$ per generation) has only very limited effects on the estimation of *M* (Appendix B and Figure S3).

The Estimation of *M* for the CNVRs Identified in HapMap Populations

A set of 4330 polymorphic CNVRs as recently identified in three HapMap populations,[2] including 60 YRI, 60 CEU, and 89 CHB+JPT, was investigated in this study. A total of 4187 (96.7%) of these CNVRs have close flanking SNPs with a genetic distance of 0.1 cM or less and were included in the additional analyses. The estimates of *M* showed a very diverse distribution in all three populations (Figure 1). Notably, the estimates of *M* are significantly correlated in all pairwise comparisons among three continental populations ($p < 10^{-5}$), especially when excluding the CNVRs with pairwise *Fst* in the largest decile (Figure S4), suggesting the consistency of the estimation of *M* among populations.

Approximate Estimation of Mutation Rates for the CNVRs Identified in HapMap Populations

Considering the observation that the statistic *M* can be affected by the demographic history of human

populations, we converted the estimates of M to the approximate mutation rate by using a simulation approach under the proper demographic model in each population and made M comparable across populations. The distribution of M according to each 1000 simulations under different demographic models[19] and at different mutation rates (5×10^{-7}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, and 5×10^{-3} per generation) is shown in Figure S5 (e.g., the mean ± SD of M is 0.258 ± 0.174 for YRI, 0.158 ± 0.157 for CEU, and 0.150 ± 0.151 for CHB+JPT under the mutation rate of 5×10^{-5} per generation; 1.10 ± 0.198 for YRI, 1.11 ± 0.152 for CEU, and 1.12 ± 0.153 for CHB+JPT under the mutation rate of 5×10^{-3} per generation). The range of mutation rates selected for simulation was based on the contention that locus-specific mutation rates for CNV are approximately 100 to 10,000 times greater than nucleotide substitution rates, i.e., ~10^{-6} to 10^{-4} per generation.[3] The range of M (2.5–97.5 percentile) for a given mutation rate in each population (Figure 1) provides an approximate estimation of the order of magnitude of the CNV mutation rate.

Similar to the findings of a wide range of variability of mutation rates among different CNV loci in the laboratory mouse strain,[5] the estimation of the CNV mutation rate in human populations also showed variances of orders of magnitude across the human genome. Among the 4187 CNVRs investigated in the estimation of M, the mutation rates for most loci were estimated at the order of 10^{-5} per generation. However, the mutation rates for CNVRs could reach the order of 10^{-3} per generation, and these CNVRs are likely mutational hotspots. A total of 104 (2.5%) potential hotspots were identified in three HapMap populations (Figure S6). All of the identified CNV mutational hotspots are listed in Table S2. Interestingly, 49 (47%) of these identified CNV hotspots involve known genes of the human genome, suggesting their implications in biological functions and human health (Table 1).

Notably, these estimators of CNV mutation rate are in concordance with experimental observations at individual loci. For example, mutation rates of two CNVRs involving the α-globin genes (chr16:162,083–167,514 and chr16:164,519–165,801, NCBI build 36) were evaluated at the order of ~10^{-5} in this study, consistent with the molecular findings in two previous studies.[7,8] In addition, the CNVR of human β-defensin genes (chr8:7,330,051–7,342,809, NCBI build 36) was suggested to be a mutational hotspot in all three populations in this study. Interestingly, this psoriasis-associated β-defensin locus[25] was previously reported among the fastest-mutating CNVs, with a germline mutation rate of 7×10^{-3} per gamete, revealed by a pedigree-based approach.[11]

Genomic Architecture and Rearrangement Mechanisms Underlying Hotspots

We also studied some genomic features of CNV loci between candidate hotspots (72 in YRI, 37 in CEU, and 22 in CHB+JPT) and nonhotspots (3075 in YRI, 2323 in CEU, and 1593 in CHB+JPT) to investigate the potential cause of a CNV hotspot. We found that about 60% of hotspots overlap with SDs, whereas only < 20% of nonhotspot CNVRs involve SDs; $p = 7.63 \times 10^{-15}$ (Fisher's exact test) for YRI, $p = 4.39 \times 10^{-12}$ for CEU, and $p = 7.62 \times 10^{-9}$ for CHB+JPT (Figure 2A).

In addition, no significant differences between hotspots and nonhotspots were observed for the recombination rate between the CNVR and its flanking SNPs; $p = 0.138$ (Mann-Whitney test) for YRI, $p = 0.971$ for CEU, and $p = 0.903$ for CHB+JPT (Figure 2B). In contrast, the CNVR sizes at the mutational hotspots are larger than those at the remaining loci; $p = 0.007$ (Mann-Whitney test) for YRI, $p = 0.042$ for CEU, and $p = 9.07 \times 10^{-5}$ for CHB+JPT (Figure 2C). The genetic diversity of the CNVR, measured by the inferred heterozygosity, is significantly higher for hotspots than that for nonhotspots as well; $p < 10^{-5}$ (Mann-Whitney test) for all of the populations (Figure 2D).

The CNV formation mechanism has been hypothesized to affect mutation rate. The nonallelic homologous recombination (NAHR) mechanism was previously proposed to be a major mechanism underlying CNV formation.[23] For example, NHAR was shown to account for > 99% of the neuropathy-associated CNVs in the 17p12 region.[26] In addition, the loci of NAHR-mediated CNVs have been regarded as "hotspots" in a previous study.[27] Here, we studied the data of CNV formation mechanisms provided by Conrad et al.[2] and found that the NAHR-mediated events were more frequent at CNV hotspots than at the remaining loci, but not significantly; $p = 0.0518$ (Fisher's exact test) for YRI, $p = 0.154$ for CEU, and $p = 0.0629$ for CHB+JPT (Figures 2E and 2F), suggesting the involvement of the NAHR mechanism in generating CNV hotspots.

These findings shed light on the understanding of genome architecture and rearrangement mechanisms underlying CNV mutations in the human genome.

Reevaluation of the Relative Mutation Rate of Deletion versus Duplication

The CNV mutation rates associated with deletion and duplication (alternatively, loss and gain) can be interrogated separately in order to investigate the respective contribution of deletion and duplication in CNV mutation. It has been proposed that deletions occur more frequently than duplications for NAHR-mediated CNVs, reflecting the lack of intrachromatid NAHR events in duplication rearrangement.[10] The deletion to duplication ratio of ~2:1 was observed consistently at three autosomal CNV loci in a pooled sperm PCR assay.[10] At the 421 NAHR-mediated CNV loci (both hotspots and nonhotspots) investigated in this study, the proportion of deletion events was estimated to be 0.774 ± 0.395 (mean ± SD) for YRI, 0.715 ± 0.436 for CEU, and 0.667 ± 0.448 for CHB+JPT. In contrast, at the 182 VNTR-mediated CNV loci, the proportion of deletion events is 0.499 ± 0.472 for YRI, 0.352 ± 0.458 for CEU, and 0.394 ± 0.474 for CHB+JPT. The proportion of deletion mutation events was significantly different between NAHR-mediated CNVRs and VNTR-mediated CNVRs (Mann-Whitney

Table 1. The Potential CNV Mutational Hotspots Involving One or More Genes

CNV ID	Chr.	Start (hg18)	End (hg18)	Location	Gene(s)	YRI		CEU		CHB+JPT	
						M	Deletion (%)	*M*	Deletion (%)	*M*	Deletion (%)
CNVR95_full	1	17548473	17551517	1p36.13	*PADI4*	0.730	0.0	0.555	26.2	0.489	61.0
CNVR116_full	1	25457812	25537782	1p36.11	*RHD*	0.671	0.5	0.714	0.0	0.445	0.0
CNVR299_full	1	109988369	110060631	1p13.3	*GSTM1, GSTM2, GSTM4, GSTM5*	0.696	7.2	0.610	26.2	0.703	62.5
CNVR299.4	1	110016535	110046454	1p13.3	*GSTM1, GSTM2, GSTM4, GSTM5*	0.757	0.0	0.399	0.0	0.379	0.0
CNVR481_full	1	204622578	204670033	1q32.1	*SRGAP2*	0.424	0.0	0.747	2.1	0.672	0.0
CNVR485.1	1	205763104	205821509	1q32.2	*CR1*	0.592	0.1	0.762	0.0	0.681	0.0
CNVR932.1	2	112761633	112766181	2q13	*ZC3H6*	0.565	0.7	0.538	0.0	0.779	0.2
CNVR2151.1	4	166377398	166378505	4q32.3	*KLHL2*	0.795	0.9	0.806	0.0	0.721	0.0
CNVR2664.1	5	159282379	159283692	5q33.3	*ADRA1B*			0.765	0.1	0.776	0.0
CNVR2728.1	6	200650	329973	6p25.3	*DUSP22*	0.403	0.3	0.718	0.0	0.802	1.2
CNVR2843.3	6	32066855	32093500	6p21.32	*C4A, C4B*	0.739	2.0	0.602	0.0	0.600	0.0
CNVR3130.1	6	160431936	160432431	6q25.3	*IGF2R*	0.629	0.0	0.755	0.0	0.588	0.0
CNVR3426.2	7	64204377	64274741	7q11.21	*INTS4*			0.833	1.7	0.807	0.0
CNVR3447.1	7	71682534	71684330	7q11.22	*TYW1B*	0.746	0.0	0.760	0.2	0.499	2.4
CNVR3561_full	7	126301909	126340192	7q31.33	*GRM8*			0.589	0.1	0.792	0.0
CNVR3585_full	7	141388076	141441024	7q34	*MGAM*	0.459	1.0	0.827	14.8	0.745	25.6
CNVR3618.4	7	151532774	151539767	7q36.1	*MLL3*	0.507	0.0	0.725	0.0	0.785	0.0
CNVR3689.1	8	584449	589454	8p23.3	*ERICH1*	0.731	96.2	0.474	96.3	0.607	100.0
CNVR3771.6	8	7330051	7342809	8p23.1	*DEFB105A, DEFB106A, DEFB107B*	0.773	0.3	0.883	3.3	0.778	4.1
CNVR4280.1	9	41552631	41641509	9p12	*ZNF658B*	0.737	0.0	0.647	0.0	0.792	0.0
CNVR4685.1	10	33229325	33230534	10p11.22	*ITGB1*	0.610	0.0	0.637	0.0	0.808	0.1
CNVR4729.3	10	51158267	51158937	10q11.23	*PARG*	0.398	2.9	0.829	0.0	0.686	0.0
CNVR5084.1	11	17166514	17167872	11p15.1	*PIK3C2A*	0.676	99.8	0.049	100.0		
CNVR5436.1	12	11917753	11918281	12p13.2	*ETV6*	0.687	100.0	0.048	100.0	0.467	100.0
CNVR6188.1	14	73064141	73121720	14q24.3	*HEATR4, ACOT1, ACOT2*	0.762	0.1	0.569	0.0	0.224	0.0
CNVR6538_full	15	97007643	97011339	15q26.3	*IGF1R*			0.770	29.7		
CNVR6668.2	16	21641198	21716960	16p12.1-p12.2	*OTOA*	0.210	0.0	0.590	2.3	0.829	0.6
CNVR6764.3	16	68702855	68754003	16q22.1	*MRCL, PDPR*	0.617	0.0	0.766	0.0	0.625	0.0
CNVR6767.2	16	69404749	69760030	16q22.2	*HYDIN, HYDIN2*	0.801	0.0	0.667	0.0		
CNVR6769_full	16	70646001	70669798	16q22.3	*DLP, HP, HPR*	0.687	0.0			0.572	0.0
CNVR6852.1	16	86580132	86581743	16q24.2	*BANP*	0.735	1.2				
CNVR6956.1	17	3988918	3989416	17p13.2	*ZZEF1*	0.703	0.0	0.026	0.0		
CNVR7021.1	17	20285952	20335955	17p11.2	*LGALS9B*	0.907	0.0	0.801	0.0	0.476	2.3
CNVR7098.1	17	36785827	36793150	17q21.1	*KRT34*	0.762	1.3	0.587	0.7	0.634	15.5
CNVR7144.1	17	53042845	53044836	17q22	*MSI2*	0.700	100.0	0.028	100.0	0.028	100.0
CNVR7154_full	17	55766887	55768922	17q23.2	*USP32*	0.710	100.0	0.017	100.0	0.021	100.0

Table 1. *Continued*

CNV ID	Chr.	Start (hg18)	End (hg18)	Location	Gene(s)	YRI		CEU		CHB+JPT	
						M	Deletion (%)	*M*	Deletion (%)	*M*	Deletion (%)
CNVR7370.1	18	62370731	62372179	18q22.1	*CDH7, CDH19*	0.829	100.0	0.077	100.0	0.012	100.0
CNVR7492_full	19	2860172	2861422	19p13.3	*ZNF57*	0.761	100.0	0.022	100.0	0.019	100.0
CNVR7540.1	19	11900599	11907033	19p13.2	*ZNF700*	0.848	3.3	0.034	0.5	0.025	0.0
CNVR7702.1	19	56823616	56842035	19q13.33	*SIGLEC5, SIGLEC14*	0.898	70.1	0.865	64.6	0.417	72.8
CNVR7708.1	19	58014801	58053170	19q13.41	*ZNF28, ZNF468*	0.686	0.0	0.665	0.0	0.618	0.2
CNVR7726.1	19	60168046	60169697	19q13.42	*NLRP2, NLRP7*	0.693	0.8	0.320	5.0	0.102	12.7
CNVR7849.1	20	41705581	41707310	20q13.12	*IFT52*	0.813	98.5	0.770	100.0	0.011	100.0
CNVR8030.1	21	43794624	43797920	21q22.3	*HSF2BP*	0.740	99.9	0.676	97.5	0.012	100.0
CNVR8066.1	21	46434207	46435262	21q22.3	*LSS*	0.716	1.7				
CNVR8067.1	21	46481946	46483100	21q22.3	*MCM3APAS, MCM3AP*	0.693	96.8	0.562	99.6	0.011	100.0
CNVR8085.1	22	16871523	16873123	22q11.21	*MICAL3*	0.542	1.0	0.796	6.4	0.782	1.5
CNVR8136.1	22	31257932	31258517	22q12.3	*SYN3*	0.394	0.6	0.752	0.2	0.623	1.6
CNVR8171_full	22	41209145	41335158	22q13.2	*SERHL, SERHL2, RRP7A, RRP7B, POLDIP3*	0.708	0.2	0.705	0.6	0.636	0.2

test: $p < 10^{-5}$ for YRI, $p < 10^{-5}$ for CEU, and $p = 2.71 \times 10^{-5}$ for CHB+JPT). These observations suggested the predicted overrepresentation of deletion versus duplication at the CNV loci mediated by the NAHR mechanism.

Discussion

Previous approaches for estimating the mutation rate for CNVs by using molecular assays or prevalence calculation are locus specific, and only a few CNV loci have ever been studied in the human genome. The statistical method of CNVMut proposed in this study makes it accessible to evaluate the CNV mutation rate across the human genome, though approximately, by using vast, readily available genotyping data of human populations.

In this method, we employed LD relationship between the CNV loci and their flanking SNPs in tracing CNV mutation events by the plausible genealogical histories reconstructed by the flanking SNPs. Because CNV alleles are not readily determined experimentally, we used the EM algorithm to infer haploid copy number with the assistance of the preinferred haplotypes of flanking SNPs. According to our evaluation, the accuracy of the CNV alleles' inference was comparable to that of the SNP haplotype inferred by the EM algorithm.[28] The error rate can be controlled by less than 5%, especially with the help of flanking SNPs (data not shown). In order to eliminate the inflated estimation of the *M* statistics caused by the uncertainty of the CNV alleles' inference as much as possible, we conducted such inference for every plausible ARG and averaged the analysis (i.e., the *M* statistics). In addition, simulations were also conducted with genotype data when converting the statistic *M* to an approximate estimation of the CNV mutation rate.

The statistic *M* was proposed to measure the minimum effective counts of CNV mutation events based on plausible ARGs constructed by flanking SNPs. The estimation of *M* was based upon an assumption of the parsimony criterion with a constant evolution rate. Any violation of this assumption, such as different changes of different copies following different mutation rates, would lead to an underestimation of the statistic *M*.

In this study, actual mutation rates were obtained by rescaling *M* with the use of coalescent simulations. The genetic diversity of a CNV locus and its flanking SNPs was simulated under the assumption of neutrality. Because this estimate does not account for purifying selection, it probably represents a lower bound on the true rate. In addition, we examined the possible effect of varying local recombination rate on mutation rate estimation for the CNVRs involving the genes of α-globin and β-defensins in YRI population and found that the estimation of CNV mutation rate for both loci was unaffected ($\sim 10^{-5}$ and $\sim 10^{-3}$ per generation, respectively), similar to those estimated under the assumption of a constant recombination rate (10^{-5} per generation). It has also been illustrated (Figure S3) that both the number of flanking SNPs (when > 10) used to construct plausible ARGs and the recombination rate between adjacent loci have limited and slight effects on the estimation of *M*. Therefore, it is proper to use 20 flanking SNPs and an assumed constant

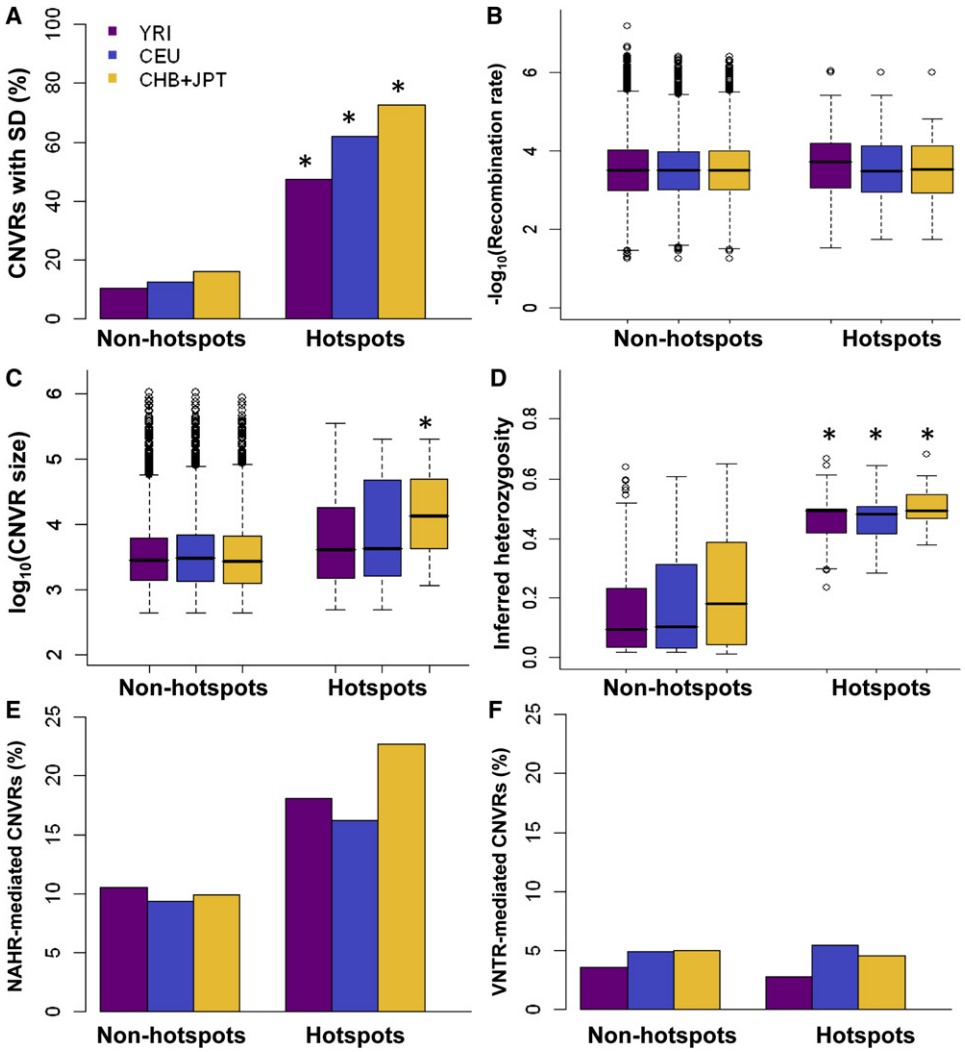

Figure 2. Comparison of CNVRs with Potential Hotspots and Nonhotspots
The percentage of CNVRs overlapping with SDs (A), the recombination rate (cM/kb) between a CNVR and its flanking SNPs (B), the CNVR size (C), the genetic diversity of the CNVR (D), the percentage of NAHR-mediated CNVRs (E), and the percentage of VNTR-mediated CNVRs (F) is illustrated for CNVRs with nonhotspots and potential hotspots. Statistically significant differences ($p < 0.001$) between the hotspot group and the nonhotspot group are indicated by asterisks.

recombination rate of 10^{-5} per generation between adjacent loci for coalescent simulations to convert the estimates of *M* into mutation rate.

Using this method, we identified 104 potential mutational hotspots for CNVRs in three HapMap populations (YRI, CEU, and CHB+JPT). The hotspots were defined as CNVRs with an approximate estimation of mutation rate at ~10^{-3} per generation. Figure S6 illustrates the consistency of 104 potential hotspots among three populations. More than one-half of the potential hotspots identified in CEU and CHB+JPT can be confirmed in another population. However, more specific hotspots were observed in YRI. It should be noted that ancient CNV mutations are more likely to be observed in Africans than in other populations, given that non-Africans experienced extreme genetic drift, especially during their migration out of Africa.

Genetic diversity for hotspots is significantly higher than that for nonhotspots (Figure 2D), which is consistent with the observation, gathered from CNVs on the human Y chromosome, that high mutation rate has driven extensive structural polymorphism on Y.[29] High genetic diversity is expected to be the consequence of high mutation rate, rather than a cause for mutational hotspots, because mutation is regarded as a force underlying genetic diversity. The enrichment of SDs at hotspots supports the contention that SDs act as mutation seeds during genome evolution and that SD-rich regions are vulnerable to genome rearrangements mediating CNVs.[23,24] Interestingly, more CNVs at hotspots than at nonhotspots are found to be mediated by the NAHR mechanism taking place between two homologous SDs (Figure 2E). It has also been shown that the other known mechanisms underlying CNV formation, including DNA repair by joining of double-strand break ends, as well as DNA replication errors, are incident to genomic regions rich in repeat and repetitive sequences,[30–33] reflecting the regional genome

architecture inciting genomic instability and mutational hotspots of CNV.

Some disease-associated loci were identified as CNV mutational hotspots in this study (Table 1). Besides the psoriasis-associated CNVs of β-defensin genes mentioned above, we also identified a CNV hotspot involving the gene of complement component *C4* (with isotopes *C4A* [MIM 120810] and *C4B* [MIM 120820]) in association with human systemic lupus erythematosus (SLE [MIM 152700]).[34] In addition, two mutational hotspots involve a glutathione conjugation gene, *GSTM1* [MIM 138350], with implication in cancer risk and drug resistance.[35] The homozygous deletion of *RHD* (MIM 111680) can cause RhD-negative blood type.[36] The CNVs at the *RHD* locus were reported to be highly polymorphic in human populations[2,37] and were also identified as CNV hotspots in this study. The *IGF1R* locus (MIM 147370) is also among CNV hotspots, and its deletion CNVs can cause gene haploinsufficiency in association with short status and other developmental defects in children born small for gestational age.[38] Notably, genomic instability was also shown in some CNVs mediating neuropsychiatric traits, including the *GRM8* [MIM 601116] and *HYDIN* [MIM 610812] loci. *GRM8* encodes the glutamatergic receptor 8, and the patient with CNV-mediated *GRM8* gene rearrangement presented autism and severe mental retardation.[39] The instable locus of hydrocephalus-associated *HYDIN* in 16q22.2 experienced an event of duplication and insertion into 1q21.1 during primate evolution, and the CNVs at both of these *HYDIN* loci can lead to abnormal brain size, i.e., microcephaly or macrocephaly.[40] These findings of functional CNV loci as mutational hotspots suggest an important role of CNV in human diseases and evolution.

In addition, though we applied this method to the published CNV genotyping data provided by Conrad et al.[2] and to genotypes of SNPs from HapMap database, various genotyping data produced by current prevailing whole-genome SNP arrays, e.g., Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina Human1M-Duo DNA Analysis BeadChip, can also be applied to this method.

In summary, we propose a statistical method, CNVMut, to achieve an approximate estimation of CNV mutation rates using genotyping data of human populations and to identify potential CNV mutational hotspots. The enrichment of repeat sequence at mutational hotspots sheds light on the understanding of genome architecture and rearrangement mechanisms underlying CNV mutations in the human genome, and the observation of functional CNV loci in hotspots implicates an important role of CNV in human health and evolution.

Appendix A: An Algorithm for Estimating CNV Mutations by Using Flanking SNPs

General Description of the Algorithm

A brief summary of the algorithm is illustrated in Figure S1 and uses an example of a CNV (0 for homozygous deletion, 1 for heterozygous deletion, 2 for homozygous wild-type) and five SNPs (0 and 1 for homozygotes, H for heterozygote) flanking the CNV. For each CNV locus, the genotypes of both the CNV and its flanking SNPs for individuals from a population are used for analysis. The ARG is constructed by using genotypes of flanking SNPs with the haplotypes inferred simultaneously (step 1). An EM algorithm is employed in determining the phase of CNV alleles with the assistance of the inferred haplotypes of flanking SNPs in step 1 (step 2). The marginal trees describing the genealogical history of each SNP are then obtained according to the ARG constructed in step 1 (step 3). Each inferred CNV allele in step 2 is then superimposed to an external node of each marginal tree as step 4 (e.g., solid and open circles for deletion and wild-type, respectively). The CNV states for the internal nodes are then inferred on the basis of the rules as defined in the subsequent section (step 5), and the number of CNV mutation events is counted (step 6). Because the real ARG is unknown, steps 1–6 are repeated R times, and a statistic M is defined as the arithmetic average of the event counts of CNV mutation in the R plausible ARGs. We have developed a program called CNVMut, which is available online, to implement the algorithm.

ARG Inference Algorithm

A method of ARG was previously proposed for reconstruction of the phylogeny on the basis of mutation, recombination, and coalescence.[15] Figure S1B illustrates an ARG of SNP haplotypes. Haplotype 2 and haplotype 4 are respectively derived from the recombination of haplotype 3 and haplotype 5. Haplotype 5 and haplotype 6 are conjunct by a mutation at site 4. When two haplotypes are identical for all SNPs, they can be conjunct by a coalescence event. For each site of the haplotype, there is a genealogical tree, called a "marginal tree," embedded in the ARG according to the impact of historical recombination events. Figure S1D illustrates the marginal trees for sites 1–3, site 4, and site 5 from left to right, respectively.

In this study, we employed the software Margarita[15] to infer ARGs and marginal trees on the basis of the genotypes of SNPs flanking a CNV. Margarita uses a heuristic algorithm to infer plausible ARGs by minimizing the number of involved recombination locally. This algorithm allows rapid computation and can handle both unphased and missing data.

Haplotype Inference Algorithm

An EM-based algorithm for inferring haplotype was proposed by Excoffier and Slatkin[41] for the analysis of SNP data. Kato et al.[16] extended the algorithm for CNV data. The difference between the algorithm for CNV and that for SNP is in the initial step during the generation of the possible haplotype pairs for a given genotype. As for the genotype of a SNP marker with its alleles coded as 0 and 1, a haplotype containing allele 0 and another containing allele 1 are the possible haplotype pair at this site. The alleles of a CNV loci are not readily distinguishable,

because the possible allele/haplotype pair for a CNV with diploid copy number of 4 can be either [0 copies/4 copies], [1 copy/3 copies], or [2 copies/2 copies] at this locus.

To superimpose CNV alleles to the marginal tree requires the inference of CNV alleles with the assistance of the pre-inferred haplotypes of flanking SNPs provided by the software Margarita. In particular, all possible haplotypes encompassing the CNV and its flanking SNPs were enumerated, and an EM algorithm was applied to calculate and update the frequencies of haplotypes in order to obtain the maximum likelihood estimation, following the iteration previously described.[16,41] In the end, the most probable haplotype pair for each individual is obtained for subsequent analysis.

Counting CNV Mutations

Each CNV allele could be superimposed to the external nodes of each marginal tree based on the inferred haplotypes, including CNV and flanking SNPs. Then, we infer the haploid copy number at each internal node of the tree. Let a binary tree unit (BTU) be denoted as a subtree with three nodes (two ends and one vertex) as shown in Figure S1D. The haploid copy number of each node is referred to as its state hereafter, and the state of each internal node is inferred according to the following rules: (1) Only when the states of both ends of the BTUs have been inferred can the state of the vertex be inferred. (2) If the state of any end is ancestral ("1" is set as the ancestral state here), the state of the vertex is set as the ancestral one. (3) If none of the states of two ends is ancestral and they are mutually exclusive, the union of the states of two ends is taken as the state of the vertex. (4) If none of the states of two ends is ancestral but not mutually exclusive, the intersection of the states of two ends is taken as the states of the vertex and both ends. (5) The procedure continues until all nodes of the marginal tree are updated.

Given that the states or haploid copy numbers of all nodes are inferred, the number of mutations can be counted for each marginal tree by traversing all of the BTUs of the marginal tree. If the states of the two ends are different, one mutation is added to the count. If L SNPs are used in ARG construction, L marginal trees and L estimators of the number of mutation are obtained. The smallest count is taken as the estimator of the plausible ARG, which can be considered as an approximate estimator of the most parsimonious mutation process. Notably, deletion mutation events and duplication mutation events can be counted separately along this parsimonious marginal tree, in order to compare mutation rates between deletion and duplication. Because the real ARG is unknown, the average of the estimator for R plausible ARGs ($R = 100$ here) is taken. For example, the statistic describing a CNV mutation can be expressed as

$$M = \frac{\sum_{r=1}^{R} \min_{l=1}^{L}\{M_{rl}\}}{NR},$$

where M_{rl} is the estimator of the mutation number for the l-th marginal tree of the r-th plausible ARG and N is the sample size.

Appendix B: Evaluation of the Performance of the Estimation of *M* with the Use of Simulated Data

Simulation Data

The program SIMCOAL2[17] was employed in simulating the neutral genetic diversity of a CNV marker and its flanking SNPs under variable mutation rates and different demographic models. Because SIMCOAL2 allows short tandem repeat (STR) markers but not CNVs, simulations were conducted by using a STR marker as a surrogate of CNV. The most common allele of STR was converted to the ancestral allele with a haploid copy number of 1. The STR alleles shorter than the common one were set as 0 for the CNV, and the STR alleles larger than the common one were set as $1 + d$ for the CNV, where d is the difference of copy number between the STR alleles and the common allele.

In the simulation, a constant population with an effective population size (*Ne*) of 5000 was assumed and the sample size was set to 100 individuals. A haploid consisting of a CNV along with 20 flanking SNPs with an MAF greater than 0.01 was generated. The mutation rate for CNV was set to 5×10^{-5} per generation, and the rate of recombination between adjacent loci (including that between the CNV locus and its flanking SNPs and that between adjacent SNPs) was 10^{-5} per generation. However, when the effect of an individual factor was interrogated, numerous values were taken in simulation. The genotype of a diploid individual was then generated by combining two randomly selected haploids.

The Estimates of *M* Increase along with Increasing Mutation Rates

The estimates of M increase along with the increasing mutation rate of CNV (5×10^{-7}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, and 5×10^{-3} per generation) for simulated data (Figure S3A), indicating that M could allow a comparison of the mutation rates among CNV loci. For example, the distributions of M are significantly ($p < 10^{-5}$; Mann-Whitney test) different between the mutation rates of 5×10^{-5} and 5×10^{-3} per generation.

Effect of Demographic Models on the Estimation of *M*

Simulation analysis showed that the estimates of M increase along with increasing effective population size, even under the same mutation rate. For example, the estimates of M increase along with the increasing effective population size (2500, 5000, and 10,000) under a given mutation rate (Figure S3B). These observations suggest that population demographic profiles must be taken into consideration when converting the estimates of M to mutation rate.

The Estimation of *M* Is Not Affected by the Number of Flanking SNPs or the Sample Size

We also investigated the possible effect of the number of flanking SNPs (10, 20, 30, 40, and 50) and the sample size (50, 100, 150, 200, and 250) on the estimation of *M*. The simulation results showed that neither an increasing number of flanking SNPs nor increasing sample size would affect the estimation of *M*, besides a slight reduction of variance (Figure S3C and S3D).

Effect of Recombination Rate between a CNV and Its Flanking SNPs on the Estimation of *M*

The LD relationship between CNVs and flanking SNPs was employed in our proposed method. The impact of recombination rate (10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, and 10^{-1} per generation) between a CNV locus and its flanking SNPs on the estimation of *M* was evaluated. Our observations indicated that the estimation of *M* was robust when the recombination rate was less than 10^{-3} per generation and that a recombination rate of 10^{-3} per generation or above can inflate the estimates of *M* (Figure S3E).

Effect of Recombination Rate between Adjacent SNPs on the Estimation of *M*

The effect of recombination rate (10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, and 10^{-3} per generation) between adjacent SNPs on the proposed algorithm for estimating *M* was also examined by simulation. It was illustrated that the estimation of *M* was robust to varying recombination rates between adjacent SNPs (Figure S3F).

Supplemental Data

Supplemental Data include six figures and two tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We are grateful to M.E. Hurles and D.F. Conrad for their kind providing of the data for CNV formation mechanisms. The project was supported by the National Natural Science Foundation of China (NSFC) (30890034 and 31000552), NSFC for the Young Distinguished Scholar (30625016), and the Program of Shanghai Subject Chief Scientist (09XD1400200). F.Z. is supported by the Shanghai Pujiang Program (10PJ1400300) and a Ministry of Education of China grant (NCET-09-0322).

Received: May 3, 2010
Revised: August 9, 2010
Accepted: September 15, 2010
Published online: October 7, 2010

Web Resources

The URLs for data presented herein are as follows:

CNVMut, http://comonca.org.cn/FDCA/resources/home_ruan.htm
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
Margarita, http://www.sanger.ac.uk/Software/analysis/margarita/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/
SIMCOAL2, http://cmpg.unibe.ch/software/simcoal2/
UCSC Genome Browser, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/

References

1. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
2. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.
3. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. Annu. Rev. Genomics Hum. Genet. *10*, 451–481.
4. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.
5. Egan, C.M., Sridhar, S., Wigler, M., and Hall, I.M. (2007). Recurrent DNA copy number variation in the laboratory mouse. Nat. Genet. *39*, 1384–1389.
6. van Ommen, G.J. (2005). Frequency of new copy number variation in humans. Nat. Genet. *37*, 333–334.
7. Lam, K.W., and Jeffreys, A.J. (2006). Processes of copy-number change in human DNA: the dynamics of alpha-globin gene deletion. Proc. Natl. Acad. Sci. USA *103*, 8921–8927.
8. Lam, K.W., and Jeffreys, A.J. (2007). Processes of de novo duplication of human alpha-globin genes. Proc. Natl. Acad. Sci. USA *104*, 10950–10955.
9. Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. Nat. Genet. ***39*** (7 Suppl), S43–S47.
10. Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat. Genet. *40*, 90–95.
11. Abu Bakar, S., Hollox, E.J., and Armour, J.A. (2009). Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins. Proc. Natl. Acad. Sci. USA *106*, 853–858.
12. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.
13. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res. *16*, 949–961.
14. Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M., and Eichler, E.E. (2006). Linkage

disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am. J. Hum. Genet. *79*, 275–290.
15. Minichiello, M.J., and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. Am. J. Hum. Genet. *79*, 910–922.
16. Kato, M., Nakamura, Y., and Tsunoda, T. (2008). An algorithm for inferring complex haplotypes in a region of copy-number variation. Am. J. Hum. Genet. *83*, 157–169.
17. Laval, G., and Excoffier, L. (2004). SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics *20*, 2485–2487.
18. Weir, B., and Cockerham, C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.
19. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.
20. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241–247.
21. Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.
22. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
23. Stankiewicz, P., and Lupski, J.R. (2002). Genome architecture, rearrangements and genomic disorders. Trends Genet. *18*, 74–82.
24. Bailey, J.A., and Eichler, E.E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. Nat. Rev. Genet. *7*, 552–564.
25. Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M., et al. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. Nat. Genet. *40*, 23–25.
26. Zhang, F., Khajavi, M., Connolly, A.M., Towne, C.F., Batish, S.D., and Lupski, J.R. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat. Genet. *41*, 849–853.
27. Itsara, A., Cooper, G.M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R.M., Myers, R.M., Ridker, P.M., Chasman, D.I., et al. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. *84*, 148–161.
28. Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. *68*, 978–989.
29. Repping, S., van Daalen, S.K., Brown, L.G., Korver, C.M., Lange, J., Marszalek, J.D., Pyntikova, T., van der Veen, F., Skaletsky, H., Page, D.C., and Rozen, S. (2006). High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat. Genet. *38*, 463–467.
30. Gu, W., Zhang, F., and Lupski, J.R. (2008). Mechanisms for human genomic rearrangements. Pathogenetics *1*, 4.
31. Argueso, J.L., Westmoreland, J., Mieczkowski, P.A., Gawel, M., Petes, T.D., and Resnick, M.A. (2008). Double-strand breaks associated with repetitive DNA can reshape the genome. Proc. Natl. Acad. Sci. USA *105*, 11845–11850.
32. Zhang, F., Carvalho, C.M., and Lupski, J.R. (2009). Complex human chromosomal and genomic rearrangements. Trends Genet. *25*, 298–307.
33. Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat. Rev. Genet. *10*, 551–564.
34. Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., et al. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am. J. Hum. Genet. *80*, 1037–1054.
35. Huang, R.S., Chen, P., Wisel, S., Duan, S., Zhang, W., Cook, E.H., Das, S., Cox, N.J., and Dolan, M.E. (2009). Population-specific GSTM1 copy number variation. Hum. Mol. Genet. *18*, 366–372.
36. Colin, Y., Chérif-Zahar, B., Le Van Kim, C., Raynal, V., Van Huffel, V., and Cartron, J.P. (1991). Genetic basis of the RhD-positive and RhD-negative blood group polymorphism as determined by Southern analysis. Blood *78*, 2747–2752.
37. Park, H., Kim, J.I., Ju, Y.S., Gokcumen, O., Mills, R.E., Kim, S., Lee, S., Suh, D., Hong, D., Kang, H.P., et al. (2010). Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat. Genet. *42*, 400–405.
38. Ester, W.A., van Duyvenvoorde, H.A., de Wit, C.C., Broekman, A.J., Ruivenkamp, C.A., Govaerts, L.C., Wit, J.M., Hokken-Koelega, A.C., and Losekoot, M. (2009). Two short children born small for gestational age with insulin-like growth factor 1 receptor haploinsufficiency illustrate the heterogeneity of its phenotype. J. Clin. Endocrinol. Metab. *94*, 4717–4727.
39. Cuscó, I., Medrano, A., Gener, B., Vilardell, M., Gallastegui, F., Villa, O., González, E., Rodríguez-Santiago, B., Vilella, E., Del Campo, M., and Pérez-Jurado, L.A. (2009). Autism-specific copy number variants further implicate the phosphatidylinositol signaling pathway and the glutamatergic synapse in the etiology of the disorder. Hum. Mol. Genet. *18*, 1795–1804.
40. Brunetti-Pierri, N., Berg, J.S., Scaglia, F., Belmont, J., Bacino, C.A., Sahoo, T., Lalani, S.R., Graham, B., Lee, B., Shinawi, M., et al. (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat. Genet. *40*, 1466–1471.
41. Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. *12*, 921–927.