Diversity of Human Copy Number Variation and Multicopy Genes

Peter H. Sudmant,1,* Jacob O. Kitzman,1* Francesca Antonacci,2 Can Alkan,3 Maika Malig,3 Anya Tsalenko,2 Nick Sampsas,2 Laurakay Bruhn,2 Jay Shendure,2 1000 Genomes Project,† Evan E. Eichler*‡‡

Copy number variants affect both disease and normal phenotypic variation, but those lying within heavily duplicated, highly identical sequence have been difficult to assay. By analyzing short-read mapping depth for 159 human genomes, we demonstrated accurate estimation of absolute copy number for duplications as small as 1.9 kilobase pairs, ranging from 0 to 48 copies. We identified 4.1 million “singularly unique nucleotide” positions informative in distinguishing specific copies and used them to genotype the copy and content of specific paralogs within highly duplicated gene families. These data identify human-specific expansions in genes associated with brain development, reveal extensive population genetic diversity, and detect signatures consistent with gene conversion in the human species. Our approach makes ~1000 genes accessible to genetic studies of disease association.

Copy number–variable genes in humans tend to map to duplicated sequences (i.e., segmental duplications) (1–4), but neither the copy number nor the locus identity of these genes can be accurately assessed with existing hybridization-based experimental methods. This bias stems from the inability to discriminate subtle differences over the full range of copy number by array comparative genomic hybridization (CGH) and a lack of informative probes for these regions in single-nucleotide polymorphism (SNP) genotyping platforms (5). This bias is more pronounced in cases where more copies of a duplicated gene are present, because multicopy integer states are difficult to resolve proportionally, resulting in a lack of understanding of the true extent of human copy number variation. As a result, the most dynamic and variable genes are frequently excluded from genome-wide association studies (6). We have developed a method to survey whole-genome shotgun sequence data to accurately assay specific duplicated genes and gene families for copy number from data generated primarily from the 1000 Genomes Project (7).

Results

Read depth can be used to predict accurately the copy number of duplicated genes within high-coverage human genomes (8). We applied this method to perform a copy number variant (CNV) analysis of 159 human genomes sequenced using the Illumina platform, including 15 high-coverage human genomes (12.3 to 43×) of diverse ethnicity and 141 low-coverage human genomes (1.5 to 7×) from the 1000 Genomes Project representing three populations (9). For each genome, reads were mapped to a repeat-masked human reference genome (build 36) with the mapqFAST aligner, which returns all possible mapping locations of a read. Read depth profiles were then constructed and corrected for biases introduced in library construction. Copy number prediction was performed by regression against a standard curve of regions of known copy (9). Using this approach, we estimated the absolute copy number genome-wide for windows of 3000 nonrepetitive bases each and generated heatmaps displaying copy number for all 159 human genomes (Fig. 1, A and B).

Because many of these genomes (table S1) were sequenced at low coverage, we tested the accuracy of our copy number predictions using three orthogonal methods of experimental validation (Fig. 2A) (9). We assessed our ability to detect simple events (gains, losses, and homologous losses) from 2270 events [median size 39.6 kilobase pairs (kbp), smallest event 1.6 kbp] across 109 common individuals. This demonstrated a concordance of 94 to 100%, dependent on the size and type of the event (fig. S20). To assess our ability to genotype multicopy number states, we performed 59 fluorescent in situ hybridization (FISH) experiments across 21 larger loci, ranging in copy from 2 to 48 (9). Of our copy number estimates, 93% (55 out of 59) were within ±1 of the number of FISH signals observed in interphase nuclei (e.g., Fig. 1A). Because FISH-based estimates are problematic when copy number exceeds 10, we designed a series of quantitative PCR (QPCR) assays to assess dynamic range response. These showed high correlation with our read depth–based estimates, with 7 out of 9 assays having a correlation coefficient of r > 0.92 (9). Read depth– and QPCR-based copy number estimates for the 1.9-kbp gene CCL3L1 across 150 individuals were concordant (r = 0.95), with both methods capturing the population-specific distributions of copy number at this locus (fig. S26).

CNV landscape diversity. We constructed genome-wide copy number maps across 159 genomes at 3-kbp resolution (Fig. 1, A and B) to assay the full extent of large-scale copy number variation among human populations. We identified 952 large CNVs greater than 50 kbp across 159 individuals (9). As expected, events of increasing size occur with progressively lower frequency. We noted that the majority of large events (55%, 522 out of 952) overlapped segmental duplications. Out of all events, 47% (452 of 952) were common, being observed in more than eight individuals (~5% of genomes). Association with segmental duplications also strongly influenced CNV frequency. Events entirely outside segmental duplications were generally rare, with 71% (390 of 546) detected in three or fewer individuals (~2% of genomes). By contrast, 91% (461 of 506) of most CNVs overlapping segmental duplications were observed in more than three individuals, consistent with recurrent variation in these dynamic regions. We identified 22 regions longer than 100 kbp that showed evidence of significant copy number differences between Asians, Europeans, and Africans [average Vst of 0.2; (table S9)] (10). These included duplications flanking the 17q21.31 MAPT locus, which has been associated with positive selection, rapid evolutionary turnover, and neurological
Analyses of the region across populations (9) revealed that although the inversion of this locus and its associated duplication are largely restricted to European-Mediterranean populations with 20% allele frequency, duplications overlapping the KIAA1267 gene were more common and definitive of populations of European origin (found in 33 out of 46 individuals at 50% allele frequency) but were nonexistent in all other ethnic groups examined (Fig. 1, B and C). Based on the extent of phased copy number change, we predicted and confirmed a short (155-kbp) duplication occurring at 20% allele frequency and a long version (205 kbp) at 30% allele frequency in Europeans (9). Distal to this locus, we identified a larger, more-copy number–polymorphic region (ranging from 2 to 6 copies). This more distal, population-stratified duplication encompassed the first 13 exons of the N-ethylmaleimide-sensitive (NSF) gene and shows increased copy number among Asians with the six-copy state occurring at ~25% frequency (13 out of 54 individuals) (Fig. 1, B and C). It is noteworthy that this gene is preferentially expressed in the human nervous system; reduction in its expression is associated with schizophrenia (12), and disruptions of its Drosophila ortholog lead to defective synaptic transmission (13).

**Fig. 1.** Landscape of human copy number variation. (A) CNV heatmap of a 734-kbp duplicated region flanking the 17q21.31 MAPT locus in 13 individuals (11 sequenced to high coverage). Read depth–based copy number (CN) estimations (3-kbp windows) are indicated by color (scale provided to the right). FISH at two separate loci validates these absolute CN predictions across five individuals (9). (B) Copy number landscape of the 17q21.31 locus across three different populations showing marked population stratification (159 genomes analyzed). A European-enriched duplication overlaps the gene KIAA1267 and is present on two haplotypes—a long form (205 kbp) and a short form (155 kbp). A 210-kbp duplication of the NSF gene ranges from two to six copies with increased copy number in Asians. For validation with array CGH, see fig. S31. (C) Copy number frequency histograms of the KIAA1267 and NSF duplications based on median read depth predict discrete copies. Duplications of the KIAA1267 locus are specific to Europeans at a frequency of 72%, 25% of Asians have six copies of NSF.
Additionally, we identified regions that are duplicated in most human individuals but previously were incorrectly classified as diploid in the reference genome, including 10 regions of >100 kbp each. We identified 173 segmentally duplicated regions for which the majority of genomes have a copy number greater than that of the reference genome. We thus have established a copy number baseline in humans, allowing more accurate genotyping from SNP microarrays. These data may be used in conjunction with single-channel intensity data from previous array CGH experiments to develop a calibration curve specifically for each region of the genome, allowing multiallelic loci to be more robustly genotyped on array-based platforms (Fig. 2A) (9). This CNV landscape diversity map may also be used to select an ideal reference genome (from the 1000 Genomes catalog) to maximize discriminatory power in array CGH studies, which will open more complex, copy number–polymorphic regions of the genome to further experimental characterization (Fig. 2B).

Gene copy number diversity and evolution. To assess the impact of copy number variation specifically on the coding portion of the genome, we genotyped the absolute copy of 25,832 individual RefSeq gene models (University of California, Santa Cruz, genome browser). As expected, our read depth assessment predicts that 99.3% of human genes are monomorphic (Fig. 3A). The gene family copy number diversity and evolution. (A) The genes most stratified by copy number in the human genome on the basis of ṼML analysis of European, African, and Asian populations. (B) Human-specific gene family expansions.
man gene models outside of segmental duplications show a median copy number of two. Limiting our set to all genes greater than 10 kbp, we find that 91.5% of human genes are fixed as diploid in all 159 humans examined. Of the copy number-variable gene families, including those within segmental duplications, we found that 80% vary between 0 and 5 copies, which suggests that extreme variation is limited to only a few gene families. We identified the 56 most variable gene families in humans (variance >3.0 among combined populations), ranging in median copy from 5 to ~368 (fig. S33 and table S7). These genes were dramatically enriched for segmental duplications [odds ratio (OR) = 311.3, P < 2.2 × 10^{-16}, Fisher’s exact test]. In 19 of these cases, no individual exhibited a copy number less than or equal to that of the reference genome, which suggests that some of these gene duplicates are underrepresented in the reference genome.

Because significant differences in allele frequency can be a signature of selection, we searched for genes with the most extreme differences in copy among the HapMap populations. These include the Yoruba people in Ibadan, Nigeria (YRI), Utah residents with ancestry from northern and western Europe (CEU), Japanese in Tokyo, Japan (JPT), and Han Chinese in Beijing, China (CHB). We used a V_q statistic (10) and identified 64 gene families [V_q > 2.0; (Fig. 3A and table S8)]. These gene families mapped almost exclusively to segmental duplications (OR = 72.0, P < 2.2 × 10^{-16}) and ranged in copy from 2 to 368 (77%, 49 out of 64 with <12 copies). However, we observed no significant correlation between V_q and copy number (ρ = 0.199), which indicated that the differences between populations were not an artifactual product of higher copy numbers. In general, the African YRI showed greater variance in stratified gene copy numbers compared with either Europeans or Asians (P < 1 × 10^{-4} and P < 1.2 × 10^{-9} after multiple-testing correction; Welch’s one-tailed t test), including genes with known stratification between populations, such as LILRA3 (14) and UGTB17 (15). One of the most stratified genes identified was CCL3LI (fig. S26), for which the importance of accurate multiallelic copy number genotyping has been recently highlighted because of conflicting reports of association with HIV susceptibility (16).

In addition, we discovered stratification among several previously uncharacterized gene families, many of which show reduced copy in Europeans and Asians compared with Africans (Fig. 3A) (9).

We characterized the evolutionary context of human gene copy number variation by comparatively analyzing short-read depth data from a gorilla, a chimpanzee, and an orangutan (9). We identified 53 gene families with increased copy number within the human lineage, 23 of which were diploid in each of the great apes, and 8 of which appear to be fixed in humans (Fig. 3B, fig. S38, and tables S10 and S11). Consistent with an origin after the human-ape divergence, the human-specific duplications and expansions displayed higher sequence identity than most duplicated genes (97.0% and 98.7%, respectively; P < 3.5 × 10^{-5}, Welch’s one-tailed t test). Human-specific duplicates include the genes GPRIN2 and SRGAP2, which have been implicated in neurite outgrowth and branching (17, 18); the brain-specific HYDIN2 gene, associated with micro and macrocephaly (19); DRD5, a dopamine D5 receptor; the GTF2I transcription factors whose deletion has been associated with visual- and sociability deficits among Williams-Beuren syndrome patients (20, 21); duplication of the SMN1 genes, at which copy increases ameliorate the severity of spinal muscular atrophy (SMA) deletions; and duplication of the CHRNA7 locus on 15q13.2 recently implicated in cases of intellectual disability and epilepsy (22, 23). We note that the fixed duplications we identified at HYDIN and GPRIN2 are not annotated in the reference and thus represent “hidden” duplications (fig. S40).

A single unique nucleotide (SUN) identifier map. Although our read depth–based genotyping provides absolute, rather than relative, copy number measurements and improved dynamic range relative to other platforms, it still lacks specificity within highly homologous segmental duplications. To discriminate among these paralogous sequences, we focused on the positions in the reference genome at which they diverge (Fig. 4A). We identified 4.07 × 10^{6} SUN positions within high-identity duplicated regions (table S2). SUNs are a distinct class of paralogous sequence variants that uniquely tag a specific paralog. By definition, the density of these markers diminishes with increasing duplication identity, yet we estimated that ~70% of duplications contain sufficient SUN density to allow paralog-specific genotyping (9). Nearby SNPs or rare variants could disrupt short-read mapping at SUN positions, which could confound copy number estimation of the tagged paralog. We examined the extent of such effects among SUNs in 12 unrelated individuals sequenced to high coverage (mean = 23.4×). Reads from these individuals were mapped to SUN positions, stringently requiring perfect 36-bp matches so as to exclude contamination between paralogous sequences. Essentially all autosomal SUNs (99.7%) were present in at least one individual, with most (84.8%) present in all 12 individuals (9). Among these, we noted a strong enrichment for SNPs (274,245 SNPs from the dbSNP version 130, OR = 1.6, P < 2.2 × 10^{-16}), which suggested that some SUNs may be misannotated as allelic variants (1, 24).

Paralog-specific copy number. We used our SUN database to complement our estimates of total copy number by developing genome-wide maps of paralog-specific copy number (psCN). At locations uniquely identified by an overlapping SUN, we counted the number of reads perfectly matching and, as before, inferred copy number using a linear model trained on regions of known copy. We validated the accuracy of our psCN estimates by specifically analyzing 383 high-confidence deletion intervals that had been detected and fully resolved by capillary sequencing on the same individuals (4). We found that 99.4% (308 out of 310) of deletions within unique regions and 93.2% (68 of 73) of deletions within duplicated regions were accurately predicted, which underscored the specificity of this approach.

Paralog-specific copy number genotyping revealed CNVs within duplicated gene families. For
example, at the complement factor H (CFH) locus, we detected deletions in close agreement with their known boundaries (>6 kbp on average, fig. S60) (4). Genotyping the resulting psCN at these intervals across all 159 samples reveals overall deletion allele frequencies for CFHR3/1 and CFHR1/4 of 29% and 4%, respectively, with elevated frequency among Africans at both deletions (25). We also identify rare genotypes, such as a reciprocal, single-copy amplification of CFHR1/4 in a single African individual (ABT) and a deletion with novel breakpoints in a single Asian individual (NA18563).

Copy number analysis from read mapping depth can potentially lead to misassignment of CNV events to the wrong copy because of cross-mapping between highly identical sequences. Using SUNs to resolve paralogous copies, we re-genotyped 406 large (>50 kbp) CNV events overlapping segmental duplications, previously called on the basis of total read mapping depth comparisons (9). We find that 60% (245 out of 406) of these regions show no signature of variation using SUNs and hence represent variation at one of the homologous loci. This “mirror” effect arises from cross-mapping between highly identical sequences but can be resolved by leveraging the SUN tags (Fig. 4B). Similarly, we found that a recently reported set of CNVs identified by CGH (2) contained a significant fraction of calls (1547 out of 4412, 35.1%) within duplicated sequences that show no evidence of paralog-specific CNV.

Paralog-specific gene family diversity: A prerequisite to understanding the function of highly duplicated gene families is the ability to genotype them for both their copy and content. Using our psCN approach, we reassessed 990 human genes partially or completely contained within segmental duplications (Fig. 5A). This analysis allowed us to distinguish two distinct classes of paralogs: 49.2% of duplicate genes which appear largely copy-invariant within the human species, and the remainder, which show extensive variation in copy number with some bias toward gain or loss.

A particularly peculiar set of variable genes are those mapping to “core duplicons” (26) which have undergone recent bursts of expansion within the human lineage, including NPIP, NBPF (Fig. 5B), and others (27). We confirmed these genes’ human-specific expansion but also observed them to be stratified among human populations (fig. S38 and Fig. 3A). We also observed more localized psCN patterns, with small patches showing reciprocal gain and loss but remaining unchanged in total copy, a signature consistent with interlocus gene conversion (i.e., unidirectional transfer of genetic information between duplicate copies) (reviewed in (28)). We validated a small number of these events and confirmed this signature at the Rhd blood group antigen genes RHD and RHCE (fig. S71) (9, 29). A scoring metric found 78 regions with a score exceeding that of RHD/RHCE with a significant excess of this signature, relative to shuffled controls ($P < 2.2 \times 10^{-16}$, one-sided Kolmogorov-Smirnov). Consistent with known examples and mechanisms of gene conversion, this was nearly exclusive to high-identity duplications (>95% ID) and was preferential to nearby tandem duplications (∼1-2 Mbp).

Discussion

We have leveraged next-generation sequence data to explore some of the most complex genetic variation in the human species and show that we can reliably predict absolute copy number without bias. More important, these data allowed us to assess the copy and content of specific duplicated genes. Nevertheless, several limitations remain. Some genes (∼30%) show too little paralogous variation or are too small to be reliably genotyped; in other cases, the paucity of markers combined with the low sequence coverage led to uncertainty in copy number accuracy (our analysis suggests that at least 8× genome-wide sequence coverage is required to achieve copy number accuracy above 97%). Finally,
we found 28 large regions with such extraordinary complexity that it is difficult to interpret the underlying pattern of genetic variation (e.g., fig. S34). Through our analysis, we identified that duplicated regions are more likely to be stratified between human populations when compared with copy number variation within unique regions of the genome. For example, 59 (92%) of the top 64 stratified gene families overlap segmental duplications (P < 2.2 × 10^{-16}). Remarkably, many of these highly polymorphic gene maps to duplications that promote recurrent rearrangements associated with intellectual disability, autism, schizophrenia and epilepsy. We hypothesize that the extreme polymorphism may contribute to genomic instability associated with disease and may predispose certain populations to different chromosomal rearrangements (30).

We have also defined the ~49% of gene duplicates that are largely invariant in copy among humans. Although this is based only on an assessment of 159 genomes from select populations, the fact that this fraction of genes remains copy number invariant in a milieu of recurrent unequal crossover suggests functional importance. Among these, we find a number of genes involved in neurological development and disease. We note that many of these duplicated genes are themselves incomplete and may represent nonprocessed pseudogenes, which may modulate the expression of the ancestral gene. The characterization of the most recently duplicated genes should facilitate identification of those that acquired new functions (neofunctionalization) versus those that have become pseudogenes or have partitioned their function among duplicate copies (31).

The ability to distinguish where copy number variation maps and where it does not within regions of high sequence identity facilitates breakpoint delineation within duplications, allowing us to refine the intervals associated with structural variants such as the recurrent, schizophrenia-associated deletion on 15q11.2 (fig. S70) (32). We anticipate that overlaying our SUN map with functional genomics data (e.g., chromatin immunoprecipitation sequencing or RNA-seq) may begin to resolve the epigenetic and expression landscape of duplicated regions and the ~1000 genes therein that have been largely inaccessible to genetic study of disease (33).

References and Notes
7. 3000 Genomes project, Nature, 28 October 2010, 10.1038/nature09534.
9. Materials and methods are available as supporting material on Science Online.

Rate of Gas Phase Association of Hydroxyl Radical and Nitrogen Dioxide

Andrew K. Mollner,1 Sivakumaran Valluvadasan,2 Lin Feng,1 Matthew K. Sprague,1 Mitchio Okumura,1,2 Daniel B. Milligan,2 William J. Bloss,2 Stanley P. Sander,1
Philip T. Martien,2 Robert A. Harley,3,4 Anne B. McCoy,3,4 William P. L. Carter2

The reaction of OH and NO2 to form gaseous nitric acid (HONO2) is among the most influential in atmospheric chemistry. Despite its importance, the rate coefficient remains poorly determined under tropospheric conditions because of difficulties in making laboratory rate measurements in air at 760 torr and uncertainties about a secondary channel producing peroxynitrous acid (HOOONO). We combined two sensitive laser spectroscopy techniques to measure the overall rate of both channels and the partitioning between them at 25°C and 760 torr. The result is a significantly more precise value of the rate constant for the HONO2 formation channel, 9.2 ±0.4) × 10^{-12} cm^3 molecule^{-1} s^{-1} (1 SD) at 760 torr of air, which lies toward the lower end of the previously established range. We demonstrate the impact of the revised value on photochemical model predictions of ozone concentrations in the Los Angeles airshed.

The gas phase three-body association reaction of hydroxyl radical (OH) with nitrogen dioxide (NO2) to form nitric acid (HONO2) (Eq. 1a)

\[ \text{OH} + \text{NO}_2 + \text{M} \rightarrow \text{HONO}_2 + \text{M} \quad (1a) \]

where M is N2 or O2 in air) is important throughout the lower atmosphere. It sequeseters reactive hydrogen (HO2, the sum of OH and peroxides) and reactive nitrogen (NO2, the sum of NO and NO) as long-lived nitric acid, shortening their atmospheric lifetimes and slowing catalytic cycles responsible for photochemically driven formation of air pollution in the troposphere and ozone depletion in the stratosphere.

The impact of the reaction in Eq. 1a is especially notable in polluted air masses, where its rate affects the formation of pollutants including ozone (O3), nitric acid, and fine particulate nitrate. This reaction leads to lower ozone pro-

1Arthur Amos Noyes Laboratory of Chemical Physics, Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA. 
2Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA.
3Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720–1710, USA.
4Department of Chemistry, The Ohio State University, Columbus, OH 43210, USA. 
5Collaborative Center for Environmental Research and Technology, University of California, Riverside, CA 92521, USA.

*To whom correspondence should be addressed. E-mail: mocz@caltech.edu (M.O.); stanley.p.sander@jpl.nasa.gov (S.P.P.); harley@ce.berkeley.edu (R.A.H.); mccoy@chemistry.ohio-state.edu (A.B.M.); wpcarter@ucr.edu (W.P.L.C.)