Mutational and selective effects on copy-number variants in the human genome

Supplementary Note

Gregory M. Cooper[1], Deborah A. Nickerson[1], and Evan E. Eichler[1,2]

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195
[2] Howard Hughes Medical Institute

GMC e-mail: coopergm@u.washington.edu, EEE e-mail: eee@u.washington.edu

Sequence and annotation data

Analyses were carried out using the human genome assembly hg17 (NCBI build35, May 2004). All sequence and annotations were obtained from the UCSC genome browser (http://genome.ucsc.edu) except where otherwise indicated. Gene annotations used correspond to the 'Known Gene' track, and include coding and UTR exons. Repeats were obtained via UCSC but generated by RepeatMasker (www.repeatmasker.org). *Alus* were identified by by using the "grep" Unix command to identify lines containing 'Alu' from the RepeatMasker annotation output.

We obtained structural variants (mostly copy-number variants; CNVs) annotated by a variety of analyses spanning multiple experimental techniques, populations, and sample sizes; details are listed in Table 1 of the main manuscript. Annotations from the Mills et al study were obtained through that paper's supplemental material (www.genome.org); the Wong et al. annotations were obtained directly from the authors (courtesy of Ze Cheng), and we used only those variants that were seen in more than 1 individual for the subsequent analysis; the Redon et al. annotations were obtained through the Database of Genomic Variants (http://projects.tcag.ca/variation). We then merged all variants > 1 kbp in size that have any amount of overlap, generating a single, non-redundant feature set with 4,131 distinct genomic regions spanning ~613 Mbp; these regions are listed in Supplementary Table 1. This merged set of 'CNVs' is used in all subsequent analyses.

Sliding window analyses

Sliding window density calculations were performed in consecutive non-overlapping 1-Mbp windows with custom perl scripts; we treated the 'overhang' at the end of each chromosome as a whole window. We merged all annotations within a given feature set with any overlap so that each base is counted at most once. We excluded Y, M, unk, and random chromosomes. GC% and N% were determined for these same windows by counting nucleotides. Features that span breakpoints between windows were split accordingly. Supplementary Table 2 lists the density values for each feature within each window across the genome, totaling 3,030 windows.

Density correlations

For these and all following analyses, we eliminated from consideration any window that contains greater than 50% missing sequence, annotated as 'N's in hg17, reducing the total number of windows to 2,852; retaining these windows inflates correlations among features due to co-occurrence of '0' densities. Relationships between feature sets were then grossly evaluated with simple linear regression models using the R programming environment (http://cran.r-project.org/), regressing the per-window CNV density against the per-window density of each listed feature (exons, segmental duplications (SDs), and *Alus*). The *p*-values reported in the first paragraph under "Distribution of copy-number variants in the human genome" correspond to the probability that the per-window densities of these feature pairs are uncorrelated genome-wide. We also regressed CNV density against exon, SD, and *Alu* density simultaneously; each of these variables remain significant at $p < 0.01$. We note that in truth these relationships are complex and non-linear, and thus these models should only be interpreted to mean that there exist strong relationships rather than that the relationships are truly linear. Also see Figures 2 and 3 for more analyses of the relationships between CNVs, genes, and SDs.

Stratification of windows into density deciles

We ranked the per-window density values for each feature set and divided the windows into 10 evenly spaced bins ('deciles'); ties in ranking were broken randomly. Note that we placed all windows with zero density into a separate bin, generating 11 total bins: 10 equal-sized bins of windows containing at least 1 nucleotide of the feature and a zero-density bin consisting of all those windows that are devoid of the feature. This was necessary since some features have a zero-density tail that comprises more than 10% of all windows. 'Rich' and 'poor' fractions of the genome (as in "the most CNV-rich fraction of the genome") correspond to the highest-density bin and the zero-density bin of the genome, respectively. The genome-wide average refers to the average density across all bins. Statistical tests differentiating 'rich' vs 'poor' or 'rich' vs 'average', were conducted using a t-test comparing the designated sets of bins to each other.

Stratification of segmental duplication annotations

For some of the analyses, we subdivided the annotations of SDs according to the percent identities of the paralogous sequence blocks (available in the 'Genomic SuperDup' track at http://genome.ucsc.edu or http://humanparalogy.gs.washington.edu). Since SDs of varying percent identities often overlap with one another, we nested this subdivision such that any base within a segmental duplication was considered to belong to the highest percent identity of all the segmental duplications to which it belongs. In this way, all bases within a duplication of 98-100% identity are parsed out first, bases within 95%-98% are parsed from the remaining nucleotides, and 90%-95% from the regions remaining after these two steps.

Allele frequency comparisons

To compare the 'allele frequencies' for variants that do or do not overlap SDs, we considered only those variants identified in the Redon et al. study via BAC-CGH; we use the 'frequency' count as supplied in the annotations in the Database of Genomic Variants (http://projects.tcag.ca/variation). We then performed a t-test comparing the frequencies for those variants that do overlap SDs versus those that do not, and also compared those variants that overlap 98%-100% identical SDs versus those that do not.

Functional biases in gene content

We used Panther analyses to test for enrichment of functional categories of genes that overlap CNVs; we considered CNVs that do not overlap SDs independently from CNVs that do overlap SDs. We identified genes that overlap these two groups using the Table Browser at UCSC (http://genome.ucsc.edu), with the 'Known Gene' track, mapping them to 'Gene Symbols' to eliminate some of the redundancy of multiple transcripts mapping to the same gene. These lists were uploaded to a published Panther web-tool (http://www.pantherdb.org/; Thomas et al. 2003). We focused on "Molecular Function" annotations, using the set of all annotated human genes as the background set, and used Bonferroni correction for multiple-testing. Both CNVs that overlap SDs and CNVs that do not overlap SDs were depleted for genes with "Molecular function unclassified" and "Ribosomal proteins". All other significant associations are enrichments and are shown in Figure 3 using the arbitrary threshold of 0.05 (after Bonferroni correction).

To check for consistency with an independent classification system, we also performed GO-term analyses using a published web tool (http://gostat.wehi.edu.au/; Beissbarth and Speed 2004). The highest-scoring enrichment categories seen for Panther were largely consistent with the GO results, although differences in the number of genes annotated to specific categories and the organizational hierarchies (GO-terms are richer and more complex), exclude direct correspondence of functional categories in some cases.

For CNVs that overlap SDs, the top-scoring GO-terms (all significant at $p < 0.001$ after correction for multiple testing) indicated enrichment for olfactory receptor activities ('sensory perception of chemical stimulus', 'neurophysiological process', 'sensory perception of smell', etc), cell adhesion activity ('cell-cell adhesion'), and the immune response ('response to pest, pathogen, or parasite', 'response to wounding'), among others. These are similar to the enrichments seen using Panther in Figure 4 (olfactory receptor enrichment can be seen indirectly with Panther in 'receptor' and 'G-protein coupled receptor').

For CNVs that do not overlap SDs, GO-terms strongly confirmed the enrichments seen using Panther. Again using a multiple testing correction and threshold of $p < 0.001$, GO-term enrichments include signaling molecules ('development-morphogenesis', 'cell-cell signaling', 'signal transduction', 'embryonic development', etc), calcium binding proteins ('calcium ion binding'), and kinase activity ('protein serine/threonine kinase activity'), among others.

References

Beissbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464-5 (2004).

Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).

Thomas, P.D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).