High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications

Tamim H. Shaikh, Xiaowu Gai, Juan C Perin, et al.

Genome Res. published online July 10, 2009
Access the most recent version at doi:10.1101/gr.083501.108

High-Resolution Mapping and Analysis of Copy Number Variations in the Human Genome: A Data Resource for Clinical and Research Applications

Tamim H. Shaikh[1,2,*], Xiaowu Gai[3,*], Juan C. Perin[3], Joseph T. Glessner[4], Hongbo Xie[3], Kevin Murphy[5], Ryan O'Hara[3], Tracy Casalunovo[4], Laura K. Conlin[1], Monica D'arcy[5], Edward C. Frackelton[4], Elizabeth A. Geiger[1], Chad Haldeman-Englert[1], Marcin Imielinski[4], Cecilia E. Kim[4], Livija Medne[1], Kiran Annaiah[4], Jonathan P. Bradfield[4], Elvira Dabaghyan[4], Andrew Eckert[4], Chioma C. Onyiah[4], Svetlana Ostapenko[3], F. George Otieno[4], Erin Santa[4], Julie L. Shaner[4], Robert Skraban[4], Ryan M. Smith[4], Josephine Elia[7,8], Elizabeth Goldmuntz[2,9], Nancy B. Spinner[1,2], Elaine H. Zackai[1,2], Rosetta M. Chiavacci[4], Robert Grundmeier[2,3,6], Eric F. Rappaport[3], Struan F.A. Grant[1,2,4], Peter S. White[2,3,5,**], Hakon Hakonarson[1,2,4,10,**]

[1] Division of Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[2] Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA

[3] Center for Biomedical Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[4] Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[5] Division of Oncology, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[6] Division of General Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[7] Department of Child and Adolescent Psychiatry The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[8] Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA, 19104, USA

[9] Division of Cardiology, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

[10] Division of Pulmonary Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA

*These authors contributed equally to this work

**Corresponding authors

Address correspondence to:

Peter S. White, Ph.D.

Children's Hospital of Philadelphia

34th St and Civic Center Blvd , Rm 1407 CHOP North

Philadelphia, PA 19104-4318

215-590-5241 (phone)

215-590-5245 (fax)

white@genome.chop.edu

Hakon Hakonarson, M.D., Ph.D.

Children's Hospital of Philadelphia

34th St and Civic Center Blvd , Rm 1216E ARC

Philadelphia, PA 19104-4318

267-426-0088 (phone)

215-590-1397 (fax)

hakonarson@chop.edu

Abstract

We present a database of copy number variations (CNVs) detected in 2,026 disease-free individuals, using high-density, SNP-based oligonucleotide microarrays. This large cohort,, comprised mainly of Caucasians (65.2%) and African-Americans (34.2%), was analyzed for CNVs in a single study using a uniform array platform and computational process. We have catalogued and characterized 54,462 individual CNVs, 77.8% of which were identified in multiple unrelated individuals. These non-unique CNVs mapped to 3,272 distinct regions of genomic variation spanning 5.9% of the genome; 51.5% of these were previously unreported, and >85% are rare. Our annotation and analysis confirmed and extended previously reported correlations between CNVs and several genomic features such as repetitive DNA elements, segmental duplications and genes. We demonstrate the utility of this data set in distinguishing CNVs with pathologic significance from normal variants. Together, this analysis and annotation provides a useful resource to assist with the assessment of CNVs in the contexts of human variation, disease susceptibility, and clinical molecular diagnostics. The CNV resource is available at: http://cnv.chop.edu.

Copy number variation (CNV) in the human genome significantly influences human diversity and predisposition to disease (Sebat et al 2004, Sharp et al. 2005, Conrad et al., 2006, Hinds et al 2006, McCarroll et al. 2006, Redon et al. 2006, Feuk et al 2006, Sebat et al. 2007, Kidd et al. 2008; Perry et al. 2008, Walsh et al. 2008). CNVs arise from genomic rearrangements, primarily due to deletion, duplication, insertion, and unbalanced translocation events. The pathogenic role of CNVs in genetic disorders has been well documented (Lupski and Stankiewicz 2005), yet the extent to which CNVs contribute to phenotypic variation and complex disease predisposition remains poorly understood. CNVs have been known to contribute to genetic disease through different mechanisms, resulting in either imbalance of gene dosage or gene disruption in most cases. In addition to their direct correlation with genetic disorders, CNVs are known to mediate phenotypic changes that can be deleterious (Feuk et al 2006, Freeman et al. 2006). Recently several studies have reported an increased burden of rare or *de novo* CNVs in complex disorders such as autism or schizophrenia as compared to normal controls, highlighting the potential pathogenicity of rare or unique CNVs (Sebat et al. 2007; Stefansson et al. 2008; Stone et al. 2008; Walsh et al. 2008; Xu et al. 2008).Thus, more thorough analysis of genomic CNVs is necessary in order to determine their role in conveying disease risk.

Several approaches have been used to examine CNVs in the genome, including array CGH and genotyping microarrays (Albertson and Pinkel 2003, Sebat et al 2004, Iafrate et al 2004, Sharp et al 2005, Redon et al 2006, Wong et al 2007). Results from over thirty studies comprising 21,000 CNVs have been reported in public repositories (Iafrate et al 2004). However, a majority of these studies have been performed on limited numbers of individuals using a variety of non-uniform technologies, reporting methods, and disease states. In addition, these data are both substantially reiterative and enriched in CNV events that are frequently observed in one or more populations. Thus, extreme care is needed in determining whether a

particular structural variant plays a role in disease susceptibility or progression. To address these challenges, we identified and characterized the constellation of CNVs observed in a large cohort of healthy children and their parents, when available. This study utilizes uniform measures to detect and assess CNVs within the context of genomic and functional annotations, as well as to demonstrate the utility of this information in assessing their impact on abnormal phenotypes. Our analysis and annotation provides a useful resource to assist with the assessment of structural variants in the contexts of human variation, disease susceptibility, and clinical molecular diagnostics.

Results

Assessment of copy number variation in 2,026 healthy individuals

DNA samples analyzed in our study were obtained from whole blood of healthy subjects routinely seen at primary care and well-child clinic practices within the Children's Hospital of Philadelphia (CHOP) Health Care Network. All samples were uniformly genotyped using the Illumina HumanHap 550 BeadChip. Genotype data were analyzed for CNVs using Illumina's BeadStudio software in combination with CNV detection methodologies developed by our group. Data from 2,026 individuals were utilized for CNV analysis, comprising 1,320 Caucasians (65.2%), 694 African Americans (34.2%), and 12 Asian Americans (0.6%). Overall, we detected a total of 54,462 CNVs, with an average of 26.9 CNVs per individual (range 4-79) (Supplementary Table 1). Collectively, these CNVs spanned 551,995,356 unique base pairs, or approximately 19.4% of the total human genome.

A majority of the CNVs detected (77.8%) were classified as non-unique CNVs as they were observed in more than one unrelated individual (Table 1). Although it is likely that some

non-unique CNVs may represent false positives due to platform-specific artifacts, a vast majority of them are hypothesized to be real as they were detected independently in more than one unrelated individual. This is supported by our experimental validation of non-unique CNVs using quantitative PCR (see below). We selected non-unique CNVs sharing at least 80% overlap in SNP content for further analysis and annotation. Mean and median sizes of non-unique CNVs were 38.3 kb and 7.2 kb, respectively. A vast majority (93.8%) of these non-unique events shared identical start and end positions with at least one additional CNV.

The remaining 22.2% of events were classified as unique CNVs since each event was detected in just one individual. The unique CNV set likely includes a rare, individual-specific variants as well as potential false positives. The unique and non-unique datasets are available for download at http://cnv.chop.edu.

We used a combination of experimental methods to provide validation for a representative set of CNVs detected in our population, including CNVs of different size classes (Table 2). Methods included cross-platform validation with the Affymetrix 6.0 array, quantitative PCR, fluorescent *in situ* hybridization (FISH), multiplex ligation-dependent probe amplification (MLPA), and comparison with reported fosmid end sequencing results (Table 2, Methods and Supplementary Methods). The array-based comparison suggested an overall validation rate of 72.7% (Table 2). For CNVs represented by >10 probes on the Illumina platform our validation rate was over 96% with a gradual decrease in validation with reducing numbers of probes. This analysis provides a conservative estimate of the true positive rate of CNVs, categorized by probe content, detected using our methods. The validation rate for non-unique CNVs, spanning 2-9 probes, as measured by quantitative PCR, was 80%. All deletions (12/12) spanning 2-9 probes validated while duplications spanning 2-9 probes had a much lower validation rate of

50%. This combined with the array-based comparison results yields a conservative false discovery rate upper bound of 50% for CNVs spanning 2-9 probes.

Generation of CNV database and web-based resource

All CNVs identified in this study are available at http://cnv.chop.edu. A database and query engine allows users to search for and sort CNVs by a variety of criteria. Results are presented in a web-based tabular format and as a set of study-wide file downloads for all CNV determinations. The CNV database can be queried for all CNVs within a selected region defined either by chromosomal coordinates or individual gene names (Figure 1). The user can visualize all CNVs within a given interval or just focus on either the non-unique or unique CNVs. Additionally, the web browser allows further classification of the CNVs by ethnicity, size, number of SNPs within, and individual variation types, which comprises duplications and both homozygous and heterozygous deletions. Resulting CNVs can be displayed in either a tabular, graphical or combined format (Figure 1 and Supplementary Figure 1). Furthermore, the 'Map it' link allows the visualization of a particular CNV in the context of all available annotations within the UCSC genome browser (http://genome.ucsc.edu), while the '*Toronto DB*' link accesses the corresponding CNV data in the Database of Genomic Variants (http://projects.tcag.ca) (Figure 1 and Supplementary Figure 1). A link for 'Downloads' of all CNV data from a given display is available at the bottom of the web page.

The contemporary Database of Genomic Variants (http://projects.tcag.ca/variation/) (DGV) serves as a valuable repository of CNVs, with over 21,000 CNVs from 31 studies represented currently. Overall comparison with this public variant set revealed that 73.1% of our non-unique CNVs overlapped with CNVs reported in DGV. In addition, the frequency of overlap increased as a function of population frequency: 54.9% of CNV blocks with <1% frequency

overlapped with DGV CNVRs, compared to 98.8% overlap with DGV for CNV blocks with frequencies >10% (Supplementary Table 2). Conversely, only 34.4% of DGV CNVs overlapped with our non-unique CNVs. Taken together, these results indicate that the CNVs we have identified are more likely to be rare events in comparison with previously reported structural variant collections. This is consistent with the notion that platform and methodological variations may contribute significantly to these differentials.

We have also examined whether the genomic distributions of various classes of structural and functional elements were correlated with the presence or absence of CNV regions. Our results extended upon previously reported correlations and are available in supplementary material under CNV analysis and annotations and in supplementary tables (3-7) and supplementary figure 3. Ethnic-specific CNV analysis was also performed for samples of Cauacasian and African ancestries, the results of which are available in the supplementary material (Suppl. Results, Suppl. Tables 8-10, Suppl. Figures 4-6).

Applications

Interpretation of CNVs. Differences in genome coverage, resolution, technologies, cohort characteristics, and CNV reporting metrics used in prior studies can prove challenging for successfully interpreting the biological significance of particular events. In comparing our results with previously reported CNVs, data from the latter often appeared to overstate the genomic extent of actual variation, as well as to underestimate variation among individuals. One typical illustration of these effects is represented by CNVs encompassing the putative tumor suppressor gene *CSMD1* (Fig. 2). Studies from DGV collectively report 49 CNVs within this gene (mean size: 347 kb; median size: 9,560 bp), including seven duplications spanning large stretches of the gene (all derived from HapMap cell lines) and an additional five CNVs predicted to disrupt one or more *CSMD1* exons (12/49, 24.5%). Interpretation of these results might lead to the conclusion that genomic alterations of this gene are frequent and do not

necessarily predispose to disease risk. However, while our CNV set identifies 507 CNVs within this region, the mean and median sizes are 7,535 bp and 3,445 bp, respectively. Moreover, only four of our CNVs (0.8%) in this region are predicted to disrupt exonic sequence, and we did not detect any of the large duplications previously reported, suggesting the possibility that these are rare variants or *in vitro* artifacts. We observed numerous additional genomic regions with CNV distributions similar to the *CSMD1*. Thus, our dataset should facilitate further delineation of the true extent of structural variation within a given genomic region, leading to improved interpretation of the biological significance of particular events.

Assessment of pathogenicity in clinical samples. A CNV dataset generated from healthy controls has the potential to be very useful in clinical applications as a comparator with CNVs identified in diseased individuals. We demonstrate the clinical utility of our CNV collection using the example of a patient with multiple congenital anomalies, including global developmental delay and brain malformations. Interestingly, 32 of 35 CNVs identified in this individual had previously been detected in our controls, many of them at frequencies greater than 1% (Table 3). Of the remaining three CNVs, two included olfactory receptor genes and were relatively small in size. The third unique CNV, the 2nd largest CNV detected (915 kb), was a deletion in 17p13 that entirely encompasses 51 genes, including several genes involved in early embryonal development. The 915 kb deletion was validated by florescence *in situ* hybridization (data not shown). Analysis of parental samples showed that while 32 of the 35 proband CNVs were found to be inherited from a parent, the 17p13 deletion was apparently *de novo*, providing support for the potential pathogenicity of this variant based solely on control CNV evidence.

To further assess the utility of our CNV database, we examined two microdeletions recently implicated in neurological disorders. A recurrent 1.5 Mb microdeletion in 15q13.3 has been associated with a recently recognized syndrome characterized by mental retardation and

seizures (Sharp et al 2008). This microdeletion contains at least six genes, including the *CHRNA7* gene that has been implicated in epilepsy (Sharp et al 2008). An assessment of our database for CNVs in the region (chr15:28723577-30192473, hg17, NCBI build 35) yielded 36 non-unique CNVs in this region, comprising 16 deletions and 20 duplications (Figure 3); 5 of these CNVs were unique (all duplications; available at http://cnv.chop.edu). Most of the control CNVs were relatively smaller and none encompassed the entire critical region implicated in the syndrome (Figure 3), except for one unique duplication encompassing the entire region (not shown). The high prevalence of this 15q13.3 microdeletion in affected individuals along with its absence in healthy controls strongly supports a role for this deletion in the etiology of the patients' phenotypes. Furthermore, duplication CNVs in controls outnumbered the deletion CNVs, were larger in size, and more frequently affected coding sequences. This may suggest that gain in copy number of genes within this region may not be as detrimental as loss due to deletion.

In sharp contrast to the above example, CNVs seen in our database contradict the genotype-phenotype correlation made between a microdeletion in 15q11.2 and a patient with neurological disorder and speech impairment (Murthy et al 2007). In this report, a ~400 kb deletion in 15q11.2 encompassing four genes was implicated in the etiology of the patient's phenotype. An assessment of our CNV set for the region (chr15:20300000-20800000, hg17, NCBI build 35) yielded 22 CNVs (both unique and non-unique), including 15 deletions and 7 duplications. Interestingly, 12 out of 22 (55%) of the control CNVs encompassed the entire critical region implicated in the syndrome (Figure 4). Although our data does not provide conclusive evidence for or against a role for this microdeletion in abnormal phenotypes, it cautions against relying strictly on assessment of disease-derived CNVs for genotype-phenotype correlations. These findings underscore the utility of our CNV dataset in clinical diagnostics.

Discussion

We present here a dataset consisting primarily of relatively rare human genomic CNVs that were derived from 2,026 healthy individuals. The generation of this resource is aimed at serving as a reference to aid in the investigation of the clinical significance of CNVs detected in disease cohorts. We believe that this will be a valuable resource to other investigators for applications in clinical diagnostics as well as in CNV enrichment and association studies for particular disease cohorts. Currently, there are several databases, including DECIPHER (https://decipher.sanger.ac.uk/) and ECARUCA (http://agserver01.azn.nl:8080/ecaruca/ecaruca.jsp), that provide cytogenetic and clinical information on disorders known to result from CNVs. We envision a pathway in which CNV data derived from clinical samples can be compared to these clinical databases, DGV and our dataset for each CNV detected. The clinical significance of CNVs detected in the sample can then be better evaluated using several criteria, including the occurrence and frequency in healthy controls, gene content, and the phenotype being studied.

Genome-wide analyses such as ours are highly dependent on the resolution and content of the discovery platform used. The platform employed in our study provides lower SNP coverage in regions of known common CNVs, regions of segmental duplication, and both the X and Y chromosomes, and as such is by no means comprehensive. Interestingly, our non-unique CNV rate was much higher than those reported in previous studies (Redon et al. 2006). The higher rate of non-unique CNVs observed in our study can be attributed at least in part to our larger study cohort. The relationship between non-unique rate and sample size approaches a plateau as more samples are surveyed (Supplementary Fig. 7), suggesting that the majority of events detectable by our methods and platform are being captured. However, recent sequence-

based analyses of CNVs, such as the fosmid end-sequencing study of nine HapMap individuals (Kidd et al 2008), indicate that a large number of as-yet undiscovered variants are present in the human genome. Thus, we conclude that although not comprehensive, our survey is identifying a substantial proportion of moderately common and rare genomic variations existing in the Caucasian and African American populations, and a considerably larger set of variants than currently exists in DGV. This observation further highlights the utility of our CNV collection for clinical applications, as moderately recurrent and rare CNVs are more likely to cause erroneous genotype-phenotype correlations.

Further, analyses such as ours are also highly dependent on computational algorithms used for detection and platform-specific experimental errors. As the large set of CNV predictions has precluded exhaustive validation, we focused validation efforts on establishing general quality guidelines for guiding users. We have used a combination of computational and experimental techniques to carefully evaluate selected CNVs. Our analyses predict low false discovery and false negative rates, especially for non-unique CNVs, deletions, and CNVs spanning four or more SNPs. Furthermore, the fact that most of our non-unique CNVs overlapped with those reported by DGV from multiple studies suggests that they represent authentic CNVs. While we have provided access to all CNV predictions, we recommend particular caution in using the unique CNV data, particularly those that are represented by fewer than 4 SNPs, where independent validation using experimental methods is advised.

Our analyses largely reiterated prior associations between genomic features and CNV distributions in a larger, more uniform sample set. The presence of ethnic-specific CNV signatures is in keeping with the demonstration of greater genomic diversity amongst individuals of African descent from HapMap data (The International HapMap project 2003, Sebat et al 2007, Frazer et al 2007). Similarly, our results confirmed that CNV distributions are positively

correlated with regions of segmental duplication (Redon et al 2006). The role of segmental duplications (SDs) in generating pathogenic chromosomal rearrangements by non-allelic homologous recombination is well documented (Lupski 2007). Our findings support a proposed model where CNV generation is promoted by close proximity to SDs (Sharp et al 2005, Redon et al 2006).

As CNV determinations continue to improve in depth, resolution and inclusion, the results will empower both biological discovery and clinical application. Greater resolution will especially be important for precisely determining the extent of each CNV, the frequency with which specific genomic regions are disrupted in healthy and disease cohorts, and the biological implications of particular variants. Based on our current CNV map, it is evident that the CNVs in public collections, which are based in part on clone-based array data, may be inflated in size consistent with other recent studies (Kidd et al 2008) . This finding is highly significant especially since use of current CNV databases in clinical applications enhances the possibility of erroneously excluding disease-causing variation in patient samples. We envision that the CNV resource described here will assist investigators performing such genomic studies on medical disorders with a genomic component.

Methods

Sample population and SNP Genotyping

Subjects were primarily recruited from the Philadelphia region through the Hospital's Health Care Network, including four primary care clinics and several group practices and outpatient practices that performed well child visits. Eligibility criteria for this study included all of the following: 1) disease-free children and parents of these children in the age range of 0-18

years of age who had high quality, genome-wide genotyping data from blood samples (defined in Supplementary Methods); 2) self-reported ethnic background; and 3) no serious underlying medical disorder, including but not limited to neurodevelopmental disorders, cancer, chromosomal abnormalities, and known metabolic or genetic disorders. Genotypes from a small set of parents of the participating children were used to assess CNV heritability patterns. All subjects and/or their parents signed an informed consent permitting the use of their genotypes and healthcare records for the study. Ancestry informative markers (AIMs) available on the HumanHap550 BeadChip (Yang et al 2005) were used to evaluate eligible subjects to determine ethnicity. Where the AIMs markers contradicted self-reported ethnicity, the AIMs marker status was used in the analysis. The cohort comprised 1,320 Caucasians, 694 African Americans and 12 Asians. This cohort contained 80 complete mother-father-child trios. Further, there were 325 mother-child, 140 father-child, 59 sibling, and 10 twin relationships confirmed by genotype concordance. The remaining 1,492 samples shared no relatedness with other samples in this data set.

Samples were assayed on the Illumina Infinium™ II HumanHap550 BeadChip (Gunderson et al 2005, Steemers et al 2006) (Illumina, San Diego, USA), as previously described in our laboratory (Hakonarson et al 2007). A total of 2,026 individuals passed all quality control (QC) measures, which included >98% SNP call rate and LRR standard deviation <0.35, and qualified for the study. The version of Illumina Infinium BeadChip is consistent for all samples in this study. The standard Illumina cluster file was used for the analysis, which is generated at Illumina by running 120 Hapmap samples, running the BeadStudio clustering algorithm, and reviewing SNPs with poor performance statistics, including call frequency, cluster separation, and Hardy Weinberg equilibrium. We reviewed this clustering in reference to our typed samples to robustly establish a reference normal diploid state for each SNP. This optimization was essential to establish the true baseline from which theta (ratio of green color corresponding to genotype) and R (intensity) are calculated into B allele frequency and Log R

Ratio values. We reviewed the raw theta and R values of each SNP in called CNV regions to ensure proper clustering of normal samples and deviation of samples with a CNV call across the region. Spurious single SNP-driven signals were rejected.

CNV detection and initial analysis

The Illumina BeadStudio 3.0 software package was used for initial CNV detection analysis. LRRs and BAFs were first exported from BeadStudio. LRR values were used as an additional sample-wide genotype quality control measure, and LRRs with a standard deviation above 0.35 were excluded from the study. In our experience, Log R Ratio standard deviation provides a robust quality metric; as demonstrated in Supplementary Figure 8, samples with LRR SDs <0.35 have similar numbers of CNVs detected with our method. Furthermore, samples with LRR SDs >0.35 had significantly higher number of detected CNVs, a majority of which are expected to be false positives resulting from background.

CNV detections were then performed for the remaining genotypes using a customized analysis workflow. Briefly, chromosomes were segmented based on LRRs using the Circular Binary Segmentation algorithm implemented in the R statistical package module DNAcopy 1.7. Default parameters were used (i.e. nperm=10000, alpha=0.01, kmax=25, nmin=200, eta=0.05, overlap=0.25, trim=0.025, undo.splits="none"). Segments were then filtered based on their average LRRs and additional devised BAF statistics:

$$\text{b2.sd} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\min\left(X_i - 0, 1 - X_i, |X_i - 0.5|\right)\right)^2}$$

$$\text{b3.sd} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\min\left(X_i - 0, 1 - X_i, |X_i - 0.67|, |X_i - 0.33|\right)\right)^2}$$

The b2.sd and b3.sd for each segment were used to measure whether the BAF pattern of a segment fits the two-copy mode better than a three-copy mode, or vice versa. The paucity of AB alleles in the segmen was used to filter for deletions. For autosomal chromosomes, the thresholds used are listed in the following table:

Type of CNV	Mean LRRs	Percentage of SNPs with BAFs between 0.6 and 0.4	Relation between b2.sd and b3.sd
Heterozygous Deletion	< -0.3	≤ 4	NA
Homozygous Deletion	< -2	NA	b2.sd ≥ b3.sd
Duplication	> 0.25	NA	b2.sd ≥ b3.sd

Different LRR cutoffs were used for the X chromosome. For males, X chromosome thresholds of -2 and 0.1 were used for hemizygous deletions and duplications, respectively. For females, X chromosome thresholds of -1.5, -0.1, and 0.6 were used for homozygous deletions, heterozygous deletions, and duplications, respectively. Female X duplications and homozygous deletions were also required to have b2.sd ≥ b3.sd. Percentage of SNPs with BAFs between 0.6 and 0.4 in the segment less or equal to 4% was a requirement for calling the segment a heterozygous deletion for females as well.

CNV validations

CNV validation was conducted by a combination of experimental methods (experimental details available in Supplementary Methods). Briefly, cross-platform validation was performed on 112 Hapmap samples to provide an unbiased assessment of the accuracy and robustness of our computational methods. Illumina HumanHap550K genotypes of these Hapmap samples were obtained from Illumina and analyzed with our computational methods. Affymetrix 6.0 genotyping data sets from these same HapMap samples were obtained from Affymetrix and analyzed for CNVs using a commercial software package (Partek Genomics Suite; Partek Incorporated, St. Louis, MO; Supplementary Table 11). Quantitative PCR was used to validate a representative sample of non-unique CNVs containing fewer than 10 SNPs (Supplementary

Table 12). Finally, CNV calls made by our method were compared to those from a HapMap individual (NA12878) that was analyzed by fosmid end sequence pairs in a recently published study by Kidd and colleagues (2008).

Data availability and access: The CNV data reported here are available at http://cnv.chop.edu. These data are also available in the Database of Genomic Variants (DGV). The individual level intensity data from the Illumina arrays will be available in dbGaP under accession phs000199.v1.p1.

Acknowledgments

This work was supported in part by NIH grant GM081519 (T.H.S), Pennsylvania Department of Health grant SAP 4100037707 (P.S.W.), a Developmental Research Award from the Cotswold Foundation (H.H. & S.F.G), funds from the David Lawrence Altschuler Chair in Genomics and Computational Biology (P.S.W.), and Institutional Awards to the Center for Applied Genomics (H.H.) and the Center for Biomedical Informatics (P.S.W.) from the Children's Hospital of Philadelphia. We would like to thank all participating subjects and families for making this study possible. Alexandre Belisle, Alejandrina Estevez, Kenya Fain, Rosalie Frechette, Alexandria Thomas, and LaShea Williams provided expert assistance with data collection and management. We also acknowledge Allen Ladd and Peter Witzleb of CHOP and Smari Kristinsson, Larus Arni Hermannsson and Asbjörn Krisbjörnsson of Raförninn ehf for informatics support. The Children's Hospital of Philadelphia Institutional Review Board has approved this study.

References

Ahn, S.J., Costa, J. & Emanuel, J.R. 1996 PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Res.* **2**4, 2623-2625

Albertson, D.G. & Pinkel, D. 2003 Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **12** Spec No 2, R145-152

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. 2006 A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet*. **38**, 75-81

Feuk, L., Carson, A.R. & Scherer, S.W. 2006 Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85-97

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., et al. 2007 A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H. et al. 2006 Copy number variation: new insights in genome diversity. *Genome Res.* **16**, 949-961

Goidts, V., Cooper, D.N., Armengol, L., Schempp, W., Conroy, J., Estivill, X., Nowak, N., Hameister, H., Kehrer-Sawatzki, H.. 2006 Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. *Hum. Genet*. **120**, 270-284.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. 2005 *Nat. Genet.* **3**7, 549-554

Hakonarson, H., Grant, S.F., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Casalunovo, T. et al. 2007 A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591-594

Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. 2006 Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet*. 38, 82-85

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., Haussler, D. 2006 The UCSC Known Genes. *Bioinformatics* **22**, 1036-1046

Iafrate A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., Lee, C. 2004 Detection of large-scale variation in the human genome. *Nat. Genet.* **36**: 949-51

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B, et al. 2008 Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64

Korbel, J.O., Urban, A.E., Grubert, F., Du, J., Royce, T.E., Starr, P., Zhong, G., Emanuel, B.S., Weissman, S.M., Snyder, M., Gerstein, M.B. 2007 Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 10110-10115

Lupski, J.R. & Stankiewicz, P. 2005 Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet.* **1**, e49

Lupski, J.R. 2007 Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43-47

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D, et al. 2008 Structural variation of chromosomes in autism spectrum disorder. *Am J Hum. Genet.* **82**, 477-488

McCarroll, S.A. Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B. et al. 2006 Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86-92

Murthy, S.K., Nguyen, A.O.H., El Shakankiry, H.M., Schouten, J.P., Al Khayat, A.I., Ridha, A., Al Ali, M.T. 2007 Detection of a novel familial deletion of four genes between BP1 and BP2 of the Prader-Willi/Angelman syndrome critical region by oligo-array CGH in a child with neurological disorder and speech impairment. *Cytogenet Genome Res* **116**:135-140

Nguyen, D.Q., Webber, C. & Ponting, C.P. 2006 Bias of selection on human copy-number variants. *PLoS Genet.* **2**, e20

Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., et al. 2008 The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**: 685-95

Pinto, D., Marshall, C., Feuk, L. & Scherer, S.W. 2007 Copy-number variation in control population cohorts. *Hum. Mol. Genet.* **16** Spec No. 2, R168-173

Redon, R. Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H. ,et al. 2006 Global variation in copy number in the human genome. *Nature* **444**, 444-454

Safran, M. Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., et al. 2003 Human gene-Centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.* **31**, 142-146

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., et al. 2004 Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-528

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., et al. 2007 Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M. et al. 2005 Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78-88

Sharp AJ, Mefford HC, Li K, Baker C, Skinner C, Stevenson RE, Schroer RJ, Novara F, De Gregori M, Ciccone R, et al 2008 A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet.* 40(3):322-8

Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R., Gunderson, K.L..2006 Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3**, 31-33

Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, et al. 2008 Large recurrent microdeletions associated with schizophrenia. *Nature* **455(7210)**:232-6

Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DH, Corvin A, Craddock NJ, et al. 2008 Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455(7210)**:237-41

The International HapMap Project. 2003 *Nature* **426**, 789-796

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., et al. 2008 Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-543

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J, Grant, S.F., Hakonarson, H., Bucan, M. 2007. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665-1674

Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., et al. 2007 A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* **80**, 91-104

Yang, N., Li, H., Criswell, L.A., Gregersen, P.K., Alarcon-Riquelme, M.E., Kittles, R., Shigeta, R. et al. 2005 Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Hum. Genet.* **118,** 382-392

Xu, B., Roos, J.L., Levy, S., van Rensburg, E.J., Gogos, J.A., Karayiorgou, M. 2008 Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**: 880-5

Figure Legends

Figure 1. Copy number variation database web portal (http;//cnv.chop.edu). This view shows the 'Combined' output of non-unique CNVs in our dataset within chromosomal 'Position' chr1:1-2,000,000. The Graphical view shows the extent and type of CNVs, het del = heterozygous deletion, dup = duplication. The CNVR is indicated and the frequency graph of the CNV Blocks is also shown. The Tabular view lists additional information for each individual CNV, including subject ethnicity, chromosomal band (Chr), sequence start and end positions, size in basepairs, type of event, and number of SNPs within (SNPs). The interface also provides links to associated CNVRs and CNV Blocks, the Database of Genomic Variants (Toronto DB), genes within or overlapping the CNV (Genes), and the UCSC browser (Map It!).

Figure 2. Comparison of CNVs detected in the current cohort with DGV CNVs within the *CSMD1* gene. Top row: Chromosome 8 genomic sequence coordinates for the *CSMD1* gene. Second row: Exonic structure of the 70-exon *CSMD1* gene. Exons are depicted as red vertical lines, and the extent of the mRNA transcript is depicted as a black horizontal line. Due to the scale of the diagram, each exon is treated as an equivalent size, and exons with short intervening sequences are drawn adjacent to each other. Third row: CNVs within the *CSMD1* gene reported in the Database of Genomic Variants. CNVs with a lighter shade of blue overlap one or more *CSMD1* exons. Bottom row: CNVs within the *CSMD1* gene reported in the current study. Numbers adjacent to two CNVs (designated by asterisks) indicate the number of instances in which that exact CNV is reported. CNVs with a lighter shade of purple overlap one or more *CSMD1* exons.

Figure 3. Copy number variation within 15q13.3. Non-unique CNVs detected in our control dataset that map within 15q13.3 (chr15:28,700,577-30,302,218, hg17, NCBI build 35) are shown as custom tracks within the UCSC web browser (http://genome.ucsc.edu/). Deletions are shown as red rectangles and duplications as blue rectangles. The CNV reported by Sharp et al. (2008) is shown as a green rectangle. The UCSC known genes and segmental duplication tracks are also shown.

Figure 4. Copy number variation within 15q11.2. Non-unique CNVs detected in our control dataset that map within 15q11.2 (chr15:20,300,000-20,800,000, hg17, NCBI build 35) are shown as custom tracks within the UCSC web browser (http://genome.ucsc.edu/). Deletions are shown as red rectangles and duplications as blue rectangles. The CNV reported by Murthy et al. (2007) is shown as a green rectangle. The UCSC known genes and segmental duplication tracks are also shown.

Tables

Table 1. Summary characteristics of non-unique CNVs.

# of SNPs		Heterozygous Deletions	Homozygous Deletions	Duplications	All events
2-3 SNPs	<1 kb	4,263	334	461	5,058
	1-10 kb	8,817	773	324	9,914
	>10 kb	1,788	308	181	2,277
	All sizes	14,868	1,415	966	17,249
4-10 SNPs	<1 kb	416	59	34	509
	1-10 kb	7,304	507	393	8,204
	>10 kb	7,149	375	1,831	9,355
	All sizes	14,869	941	2,258	18,068
>10 SNPs	<1 kb	0	0	0	0
	1-10 kb	210	3	3	216
	>10 kb	4,472	215	2,165	6,852
	All sizes	4,682	218	2,168	7,068
Total	<1 kb	4,679 (84.0%)	393 (7.1%)	495 (8.9%)	5,567
	1-10 kb	16,331 (89.1%)	1,283 (7.0%)	720 (3.9%)	18,334
	>10 kb	13,409 (72.5%)	898 (4.9%)	4,177 (22.6%)	18,484
	All sizes	34,419 (81.2%)	2,574 (6.1%)	5,392 (12.7%)	42,385

Size (bp)	Heterozygous Deletions	Homozygous Deletions	Duplications	All events
Minimum	2	2	12	2
Median	5,994	3,974	44,762	7,229
Mean	26,602	13,638	125,105	38,346
Maximum	2,632,254	217,125	2,478,824	2,632,254

# of SNPs	Heterozygous Deletions	Homozygous Deletions	Duplications	All events
Minimum	2	2	2	2
Median	4	3	8	4
Mean	6.5	5.0	16.8	7.7
Maximum	170	50	524	524

Table 2. Validation of CNVs

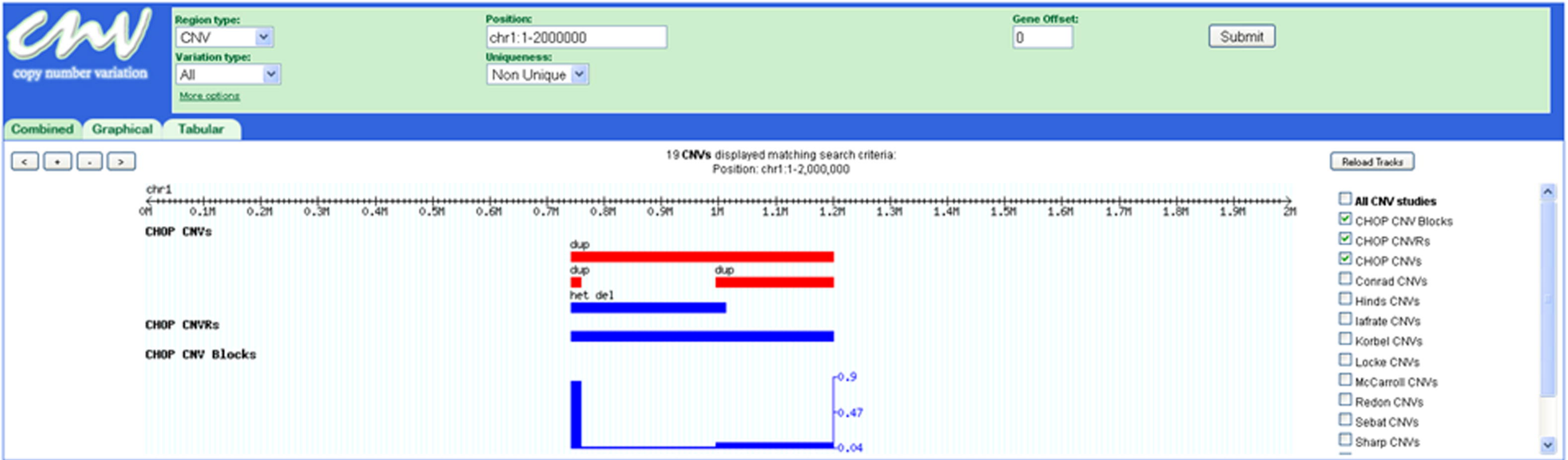
Type of CNV	Number of SNPs within CNV (Illumina array)	Experimental Technique used for validation		
		qPCR Percent validated (*)	Array-based comparison Affy vs Illumina Percent validated (*)	Comparative validation (Kidd et al 2008) Percent validated (*)
Deletion	2-3	100 (5/5)	51.3 (451/879)	89 (8/9)
	4-9	100 (7/7)	78.8 (713/905)	100 (6/6)
	>=10	NT	97.1 (330/340)	83.3 (5/6)
	All deletions	100 (12/12)	70.3 (1494/2124)	90.5 (19/21)
Duplication	2-3	0 (0/2)	78.2 (18/23)	NA
	4-9	67 (4/6)	85.7 (96/112)	NA
	>=10	NT	96 (144/150)	NA
	All duplications	50 (4/8)	90.5 (258/285)	NA
All (Total)		80 (16/20)	72.7 (1753/2409)	NA

NT = not tested; NA= not applicable; *= total # validated/total # tested

Table 3. Assessment of CNVs detected in a patient with multiple congenital anomalies.

Chr.	Sequence Start	Sequence End	# of SNPs	CNV Type	CNV Size	Gene Content*	Control CNVs (%)	P.T.	Assessment
2	3706747	3715513	3	Del	8766	None	146 (7.2%)	Yes	Variant
2	49447507	49449446	3	Del	1939	None	7 (0.3%)	Yes	Variant
3	65166887	65187636	11	Del	20749	None	162 (8.0%)	Yes	Variant
3	163613393	163625177	6	Del	11784	None	146 (7.2)	Yes	Variant
3	192548623	192552686	5	Del	4063	CCDC50	448 (22.1)	Yes	Variant
4	10073596	10076425	4	Del	2829	None	167 (8.2)	Yes	Variant
4	55498098	55499119	2	Del	1021	None	0 (0.0)	Yes	Variant
4	56028303	56029215	2	Del	912	None	0 (0.0)	Yes	Variant
4	87336261	87337106	4	Del	845	MAPK10	91 (4.5)	Yes	Variant
4	157663708	157664385	3	Del	677	None	33 (1.6)	Yes	Variant
5	99552162	99636755	13	Dup	84593	None	0 (0.0)	Yes	Variant
6	19154709	19156752	2	Del	2043	None	36 (1.8)	Yes	Variant
6	29464310	29476000	7	Del	11690	OR gene	0 (0.0)	No	Non-pathogeni
6	31349106	31350153	2	Del	1047	None	11 (0.5)	Yes	Variant
6	31385967	31416609	6	Del	30642	None	9 (0.4)	Yes	Variant
6	79031111	79088461	24	Del	57350	None	691 (34.1)	Yes	Variant
7	89165554	89169524	2	Del	3970	None	49 (2.4)	Yes	Variant
8	4460541	4472656	19	Del	12115	CSMD1	7 (0.3)	Yes	Variant
8	17625071	17625980	3	Del	909	MTUS1	75 (3.7)	Yes	Variant
8	72378670	72379585	3	Del	915	EYA1	24 (1.2)	Yes	Variant
10	20890630	20894603	6	Del	3973	None	149 (7.4)	Yes	Variant
10	126675334	126681170	6	Dup	5836	CTBP2	0 (0.0)	Yes	Variant
11	55127597	55193702	8	Dup	66105	2 OR genes	4 (0.2)	Yes	Variant
11	126556731	126564157	7	Del	7426	None	1 (0.05)	Yes	Variant
12	2115897	2120329	4	Del	4432	CACNA1C	64 (3.2)	Yes	Variant
14	85357100	85376726	5	Del	19626	None	54 (2.7)	Yes	Variant
15	18421386	19852603	6	Del	1431217	SD region	6 (0.3)	Yes	Variant
15	21948655	21948712	2	Del	57	None	1 (0.05)	Yes	Variant
15	32505886	32549650	8	Del	43764	None	55 (2.7)	Yes	Variant
17	**7142405**	**8057840**	**182**	**Del**	**915435**	**50 genes**	**0 (0.0)**	**No**	**Pathogenic**
17	21480206	22166482	39	Dup	686276	SD region	2 (0.1)	Yes	Variant
17	53561087	53598859	12	Del	37772	OR gene	0 (0.0)	No	Non-pathogeni
18	897521	897710	2	Dup	189	ADCYAP1	0 (0.0)	Yes	Variant
18	56251779	56271952	12	Del	20173	None	1 (0.05)	Yes	Variant
22	15628953	15630785	2	Del	1832	None	1 (0.05)	Yes	Variant

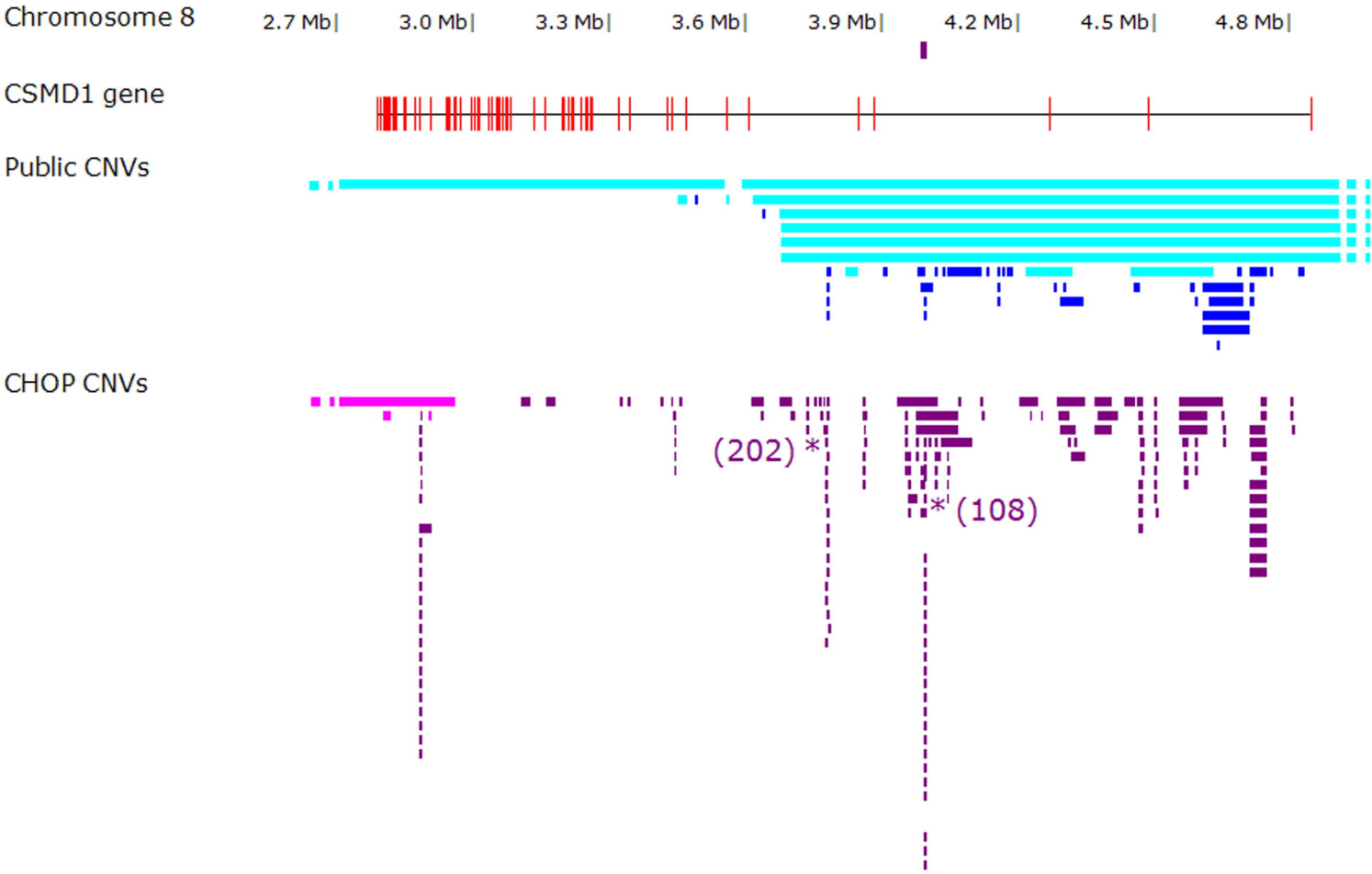
* OR gene: Olfactory receptor gene; SD region: region of known segmental duplication (RefSeq gene transcript overlap was used for gene assessment); P.T. = Parental Transmission.



19 CNVs displayed matching search criteria:
Position: chr1:1-2,000,000

Download: Download ?

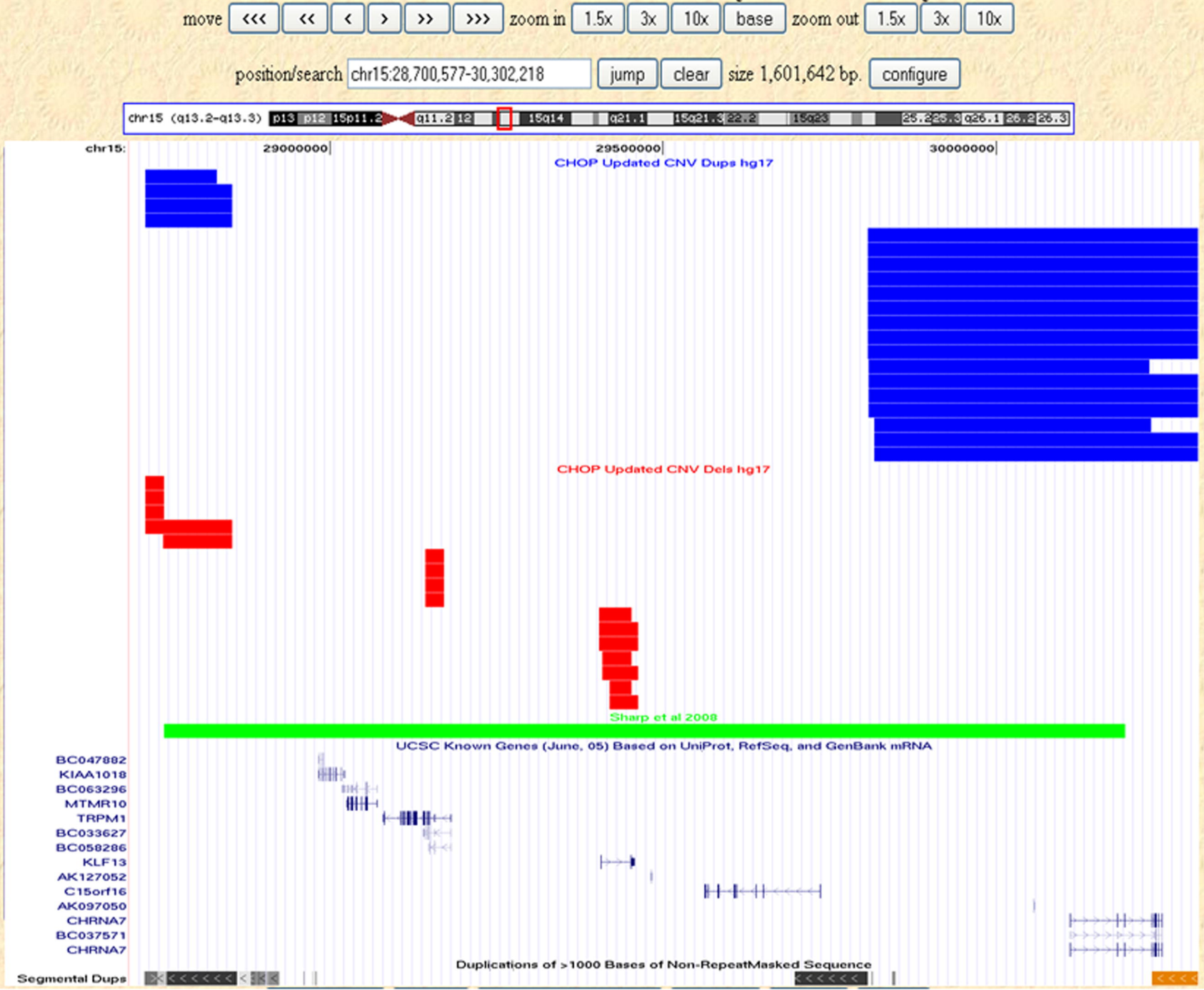
Ethnicity ▸	Chr	Start ▾	End ▸	Size ▸	Type ▸	SNPs ▸	CNVR	CNV Blocks	Toronto DB	Genes	Browser
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
African	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
African	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
African	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	758311	15883	dup	2	1	1	Toronto DB	1 gene	Map it!
Caucasian	1p36.33	742429	1011278	268850	het del	5	1	2	Toronto DB	9 genes	Map it!
African	1p36.33	742429	1201155	458727	dup	34	1	3	Toronto DB	18 genes	Map it!
African	1p36.33	995669	1201155	205487	dup	32	1	1	Toronto DB	10 genes	Map it!


Chromosome 8
2.7 Mb
3.0 Mb
3.3 Mb
3.6 Mb
3.9 Mb
4.2 Mb
4.5 Mb
4.8 Mb
CSMD1 gene
Public CNVs
CHOP CNVs
(202) *
* (108)

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr15:28,700,577-30,302,218 jump clear size 1,601,642 bp. configure



UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr15:20,300,000-20,800,000 jump clear size 500,001 bp. configure