Hum Genet
DOI 10.1007/s00439-007-0404-5

ORIGINAL INVESTIGATION

Germ-line DNA copy number variation frequencies in a large North American population

George Zogopoulos · Kevin C. H. Ha · Faisal Naqib · Sara Moore · Hyeja Kim · Alexandre Montpetit · Frederick Robidoux · Philippe Laflamme · Michelle Cotterchio · Celia Greenwood · Stephen W. Scherer · Brent Zanke · Thomas J. Hudson · Gary D. Bader · Steven Gallinger

Received: 25 May 2007 / Accepted: 9 July 2007

Abstract Genomic copy number variation (CNV) is a recently identified form of global genetic variation in the human genome. The Affymetrix GeneChip 100 and 500 K SNP genotyping platforms were used to perform a large-scale population-based study of CNV frequency. We constructed a genomic map of 578 CNV regions, covering approximately 220 Mb (7.3%) of the human genome, identifying 183 previously unknown intervals. Copy number changes were observed to occur infrequently (<1%) in the majority (>93%) of these genomic regions, but encompass hundreds of genes and disease loci. This North American population-based map will be a useful resource for future genetic studies.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-007-0404-5) contains supplementary material, which is available to authorized users.

Introduction

Copy number variation (CNV) is a well-established cause of rare genomic disorders (Freeman et al. 2006), but the presence of wide-spread CNVs in apparently phenotypically normal individuals was not recognized until recently (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005; Conrad et al. 2006; Freeman et al. 2006; Hinds et al. 2006; McCarroll et al. 2006; Redon et al. 2006; Wong et al. 2007). The presence of large structural variants in the human genome challenges the dogma that a germline diploid state represents normal copy number for all DNA regions across the entire genome.

G. Zogopoulos · K. C. H. Ha · F. Naqib · S. Moore · H. Kim · S. Gallinger
Sam Minuk Cancer Genetics and Biomarker Laboratories, Fred Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Toronto, Canada

G. Zogopoulos · S. Gallinger
Dr. Zane Cohen Digestive Diseases Clinical Research Centre, Mount Sinai Hospital, Toronto, Canada

K. C. H. Ha · F. Naqib · G. D. Bader
Banting and Best Department of Medical Research and the Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada

A. Montpetit · F. Robidoux · P. Laflamme · T. J. Hudson
The McGill University and Genome Québec Innovation Centre, Montreal, QC, Canada

M. Cotterchio · B. Zanke · S. Gallinger
Cancer Care Ontario, Toronto, Canada

C. Greenwood
Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Canada

S. W. Scherer
The Centre for Applied Genomics, The Hospital for Sick Children and Department of Molecular and Medical Genetics, University of Toronto, Toronto, Canada

B. Zanke · T. J. Hudson
Ontario Institute for Cancer Research, Toronto, ON, Canada

G. D. Bader
Rm 1225, Mount Sinai Hospital, 600 University Ave, M5G 1X5 Toronto, ON, Canada

S. Gallinger (✉)
Samuel Lunenfeld Research Institute, Toronto, Canada
e-mail: sgallinger@mtsinai.on.ca

Although a significant fraction of these germ-line deletions and gains are likely benign variants, CNVs affecting dosage-sensitive genes or regulatory regions may mediate or predispose to phenotypic outcomes. Characterization of large populations will help uncover the biological significance of CNVs and facilitate the identification of infrequent hereditary and de novo CNVs that may underlie Mendelian disease, genomic disorders, and common diseases. Therefore, to ascertain the relative contribution of CNVs to human variation and to enable the study of their role in disease predisposition, we have constructed the first North American population-based resource of copy number variable regions (CNVRs: stretches of overlapping or adjacent gains or losses of DNA) and measured the population frequencies of CNVs in these regions by characterizing 1,190 controls from Ontario, Canada.

Materials and methods

DNA samples

Biospecimens were obtained from the Ontario Familial Colorectal Cancer Registry (OFCCR), a member of the National Cancer Institute Cooperative Family Registries for Colorectal Cancer Studies (http://www.epi.grants.cancer.gov/CFR/about_colon.html) (Cotterchio et al. 2000). Approximately, 1,200 control subjects (random sample of men and women) living in Ontario, Canada were recruited by telephone from a list of randomly selected residential telephone numbers for Ontario and from population-based Tax Assessment Rolls of the Ontario Ministry of Finance. The 1,190 subjects averaged 63.5 ± 8.6 years in age, and consisted of 662 males and 516 females (data unavailable for 12 individuals). The ancestry of these subjects was self-reported as follows: Caucasian (n = 1,062), Black (n = 8), East Asian (n = 20), South Asian (n = 15), and Hispanic (n = 1). Ethnicity data were not available for 78 individuals and three subjects were of mixed ancestral backgrounds.

Study subjects donated a venous blood sample and peripheral blood lymphocytes were isolated using Ficoll-Paque, according to the manufacturer's recommendations (Amersham Biosciences, Baie d'Urfé, Quebec, PQ, Canada). Genomic DNA was extracted from lymphocytes by phenol–chloroform. The Research Ethics Board at Mount Sinai Hospital, Toronto, approved study protocols and informed consent was obtained from all enrolled study subjects.

Hybridization of samples onto the Affymetrix GeneChip Human Mapping 100 K Array Set

For target preparation prior to hybridization, 250 ng of genomic DNA were digested with either HindIII or XbaI (New England Biolabs, Ipswich, MA, USA), followed by ligation with HindIII or XbaI specific adapters (Affymetrix Inc., Santa Clara, CA, USA). The ligated DNA was diluted 4 fold, and then PCR amplified using a primer designed for the adapter DNA. The PCR reactions were purified using a Qiagen MinElute 96 UF PCR Purification Plate (Qiagen Inc., Valencia, CA, USA), and 40 µg of this purified product was fragmented using 0.2 units of DNase I (Affymetrix Inc.). The fragmented DNA was labeled using 0.214 mM DNA Labeling Reagent (DLR) (Affymetrix Inc.) and 105 U of terminal deoxy-nucleotidyl transferase (Tdt) (Affymetrix Inc.) for 2 h at 37°C. Hybridization onto the 50 K HindIII and 50 K XbaI arrays and subsequent steps were performed as described by the manufacturer (http://www.affymetrix.com).

Hybridization of samples onto the Affymetrix GeneChip Human Mapping 500 K Array Set

For target preparation prior to hybridization, 250 ng of genomic DNA were digested with either NspI or StyI (New England Biolabs), followed by ligation with NspI or StyI specific adapters (Affymetrix Inc.). The ligated DNA was diluted fourfold, and then PCR amplified using a primer designed for the adapter DNA. The PCR reactions were purified using a DNA Amplification Clean-Up Kit (Clontech, Mountain View, CA, USA), and 90 µg of this purified product was fragmented using 0.25 units of DNase I (Affymetrix Inc.). The fragmented DNA was labeled using 0.857 mM DNA Labeling Reagent (DLR) (Affymetrix Inc.) and 105 U of terminal deoxy-nucleotidyl transferase (Tdt) (Affymetrix Inc.) for 4 h at 37°C. Hybridization onto the 250 K NspI and 250 K StyI arrays and subsequent steps were performed as described by the manufacturer (http://www.affymetrix.com).

CNVR determinations

To identify a gain or loss in genomic copy number, we used the Copy Number Analyzer for GeneChip (CNAG) algorithm (Nannya et al. 2005), which was developed specifically for measuring copy number alterations using the Affymetrix GeneChip 100 and 500 K platforms. In proof-of-principle experiments, we used this algorithm to accurately predict a known structural genomic change (deletion of exons 1–16 in the *MSH2* gene) in a previously characterized clinical sample, and to identify, in control samples, previously published CNVs. We then proceeded with our population-based study, analyzing each chip separately and pooling the results from all four chips across all 1,190 samples to delineate CNVRs.

We implemented several filter criteria in our analyses of Affymetrix GeneChip data to ensure high quality of the

resultant CNVRs. Hybridization experiments with genotyping call rates of <93%, using the Dynamic Model (100 K platform) (Di et al. 2005) or BRLMM (500 K platform) (Hua et al. 2007) algorithms, were not included in the analysis. CNAG version 2 software (http://www.genome.umin.jp), which employs a Hidden Markov Model, was used to identify markers (probes) showing copy number variation greater or less than diploid (Nannya et al. 2005). The CNAG algorithm includes a built-in correction for signal-to-noise ratios, which uses quadratic regressions to account for the GC content and the length of the PCR products. Hybridization data from each Affymetrix chip (HindIII, XbaI, NspI and StyI) were analyzed separately. For each chip analysis, the sample set was divided equally and randomly into test and reference sets. The algorithm requires reference DNA sources to measure copy number alterations. Following the analysis of the first set of test samples, the data sets were swapped and the CNAG algorithm was rerun. In this manner, all samples were screened for copy number changes.

Prior to determining the CNVRs, a set of stringently defined filters was computationally applied on all preliminary copy number data. In order to reduce noise in the copy number data generated by CNAG, small 1 bp singletons (copy number changes based on a single marker) were excluded from the analyses. Large chromosomal segmental imbalances spanning more than 14 Mb were rare events, likely representing somatic changes in lymphocytes and were also excluded from further analysis. Once these two filters were applied, we further restricted our analysis to samples showing copy number changes at <1000 and <200 chromosomal (SNP or markers) positions for the 100K (HindIII, XbaI) and 500K (NspI, StyI) platforms, respectively. Samples with copy number alterations at chromosomal positions exceeding these thresholds had high noise to signal ratios. In addition, only markers observed as gains (or losses in the deleted regions) in at least two separate individuals were used to define CNVRs. This filter further increased the quality of CNVR determinations and excluded very rare structural variations (population frequency <0.17%) from our map of the more common CNVs present in the Ontario population. Once these filters were applied, all markers, from all four chips showing copy number changes, were merged and CNVRs were determined. The filtering criteria resulted in the inclusion of 1,190 samples in the analyses, with 939 and 604 samples having adequate quality copy number data across 3 and 4 chips, respectively (Table 1).

Since it is not possible to determine the exact boundaries of individual copy number changes using genome-wide SNP genotyping platforms, we approximated the boundaries of genomic regions with CNVRs. Each CNVR represents the union of all observed copy number changes at multiple chromosomal positions (i.e., at the SNP or marker locations) in a stretch of genomic DNA. Setting a threshold for a maximum distance between two adjacent markers in a CNVR, permitted approximation of CNVR breakpoints, such that a breakpoint occurred when this preset inter-marker distance was exceeded. To select the appropriate threshold value, we evaluated the effect of varying this variable on the number of resultant CNVRs. Since the slope of this curve markedly changed at an inter-marker distance of approximately 1 Mb, we selected 1 Mb as the threshold distance.

Quantitative PCR

Lymphocyte genomic DNA was assayed using the 7700 ABI real-time instrument (Applied Biosystems, Foster City, CA, USA) and the Platinum SYBR Green qPCR SuperMix-UDG assay system (Invitrogen Canada Inc., Burlington, ON, Canada), according to the manufacturer's specifications. The comparative C_T method (User Bulletin #2; Applied Biosystems) was used to identify copy number changes. For each sample and primer set, triplicate reactions were run. Since copy number changes at the MSH2 locus are rare events, which have only been observed in individuals with hereditary non-polyposis colorectal cancer, we selected a region within exon 13 of the MSH2 gene as the diploid reference threshold for the qPCR assay. MSH2 gene deletions had already been excluded in all relevant OFCCR cases by a multiple ligation-dependent probe amplification assay. For each CNVR validation experiment, we selected the negative control sample (predicted by our computational approach to be diploid at the CNVR locus) with a ΔC_T value closest to zero and used it as the diploid reference for that CNVR. A *t*-test comparing ΔC_T values was used to determine the statistical significance of the result. A significant change in copy number was observed when $P < 0.05$. Primer sets and PCR conditions are available upon request.

Table 1 Details of samples satisfying the filtering criteria

Samples (*n*=)	100K platform		500K platform		4 chips per sample (*n*=)	3 chips per sample (*n*=)	Median number of chips per sample (*n*=)
	HindIII (*n*=)	XbaI (*n*=)	NspI (*n*=)	StyI (*n*=)			
1190	979	978	1123	803	604	939	4

CNVR genomic feature annotation

All genomic feature information was obtained from the 'Known Genes' data track, downloaded from the University of California, Santa Cruz (UCSC) Genome Browser, May 2004 Assembly (NCBI Build 35), which corresponds to our Affymetrix chip SNP coordinates (Kent et al. 2002; Karolchik et al. 2003, 2004). Genome feature overlap was considered if at least 1 bp from the feature overlapped our CNVRs. Genes were defined by transcription start and end position.

Results

We constructed a CNVR frequency map by assaying lymphocyte genomic DNA, using the genome-wide Affymetrix GeneChip 100 and 500 K SNP genotyping platforms to identify genomic copy number changes. The use of four chips (i.e., the Affymetrix HindIII, XbaI, NspI and StyI chips) provides overlapping genomic coverage, resulting in both complementary and confirmatory CNVR results.

Using this computational approach, we assembled a genomic map consisting of 578 stringently defined CNVRs covering approximately 220 Mb (7.3%) of the human genome, with an average length of 408 kb (Fig. 1; Table 2). We found, on average, 6 ± 3 (SD) genomic regions with copy number alterations in each subject. The population frequencies of copy number changes within each CNVR were also estimated. Supplementary Table 1 provides the 578 CNVR coordinates and the population frequency and ancestry associations for each CNVR. Each CNVR consists of multiple markers (average = 22, range 2–292). For each CNVR, we counted the number of times a copy number change (gains and deletions counted separately) occurred at each marker position in our series of 1,190 samples. An average (±SD) was taken across all markers found in the CNVR and reported as the "Average Number of Samples

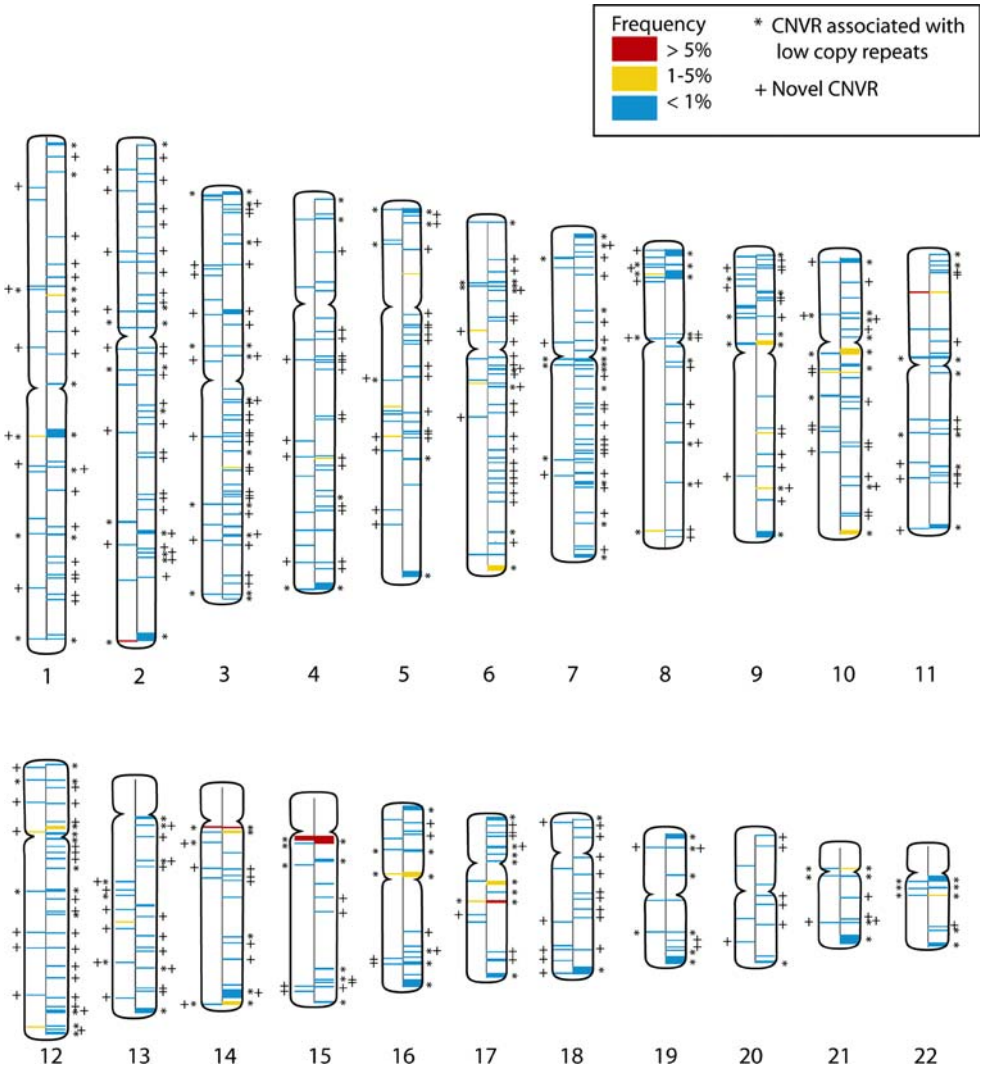

Fig. 1 Genomic map of common CNVRs found in 1,190 population-based controls in Ontario, Canada. Chromosomal locations of copy number losses and gains are shown on the left and right of the ideograms. Colors indicate whether each CNVR occurs at a population frequency of <1% (*blue*), 1–5% (*yellow*) or >5% (*red*). CNVRs associated with low copy repeats are shown (*). Novel CNVRs that were not detected in the HapMap sample collection[9] are also indicated (+)

Table 2 Characteristics of 578 CNVRs identified in 1,190 control subjects

Total # CNVRs	578
#CNVRs with gains	405 (70.1%)
#CNVRs with deletions	169 (29.2%)
#CNVRs with gains and deletions	4 (0.7%)
Average CNVR length (bp)	408, 377
Maximum CNVR length (bp)	4, 589, 135
Minimum CNVR length (bp)	12
#Informative markers across the genome[a]	12,001
Average markers per CNVR	22
Maximum markers per CNVR	292
Minimum markers per CNVR	2
#CNVRs associated with known genes	323/578 (55.9%)
#CNVRs with gains associated with known genes	245/405 (60.49%)
#CNVRs with deletions associated with known genes	75/169 (44.38%)
#CNVRs with gains and deletions associated with known genes	3/4 (75%)
Segmental duplications associated with CNVRs	195/578 (33.74%)

[a] Informative markers include all SNP markers that showed a copy number gain (or loss) in, at least, two samples

with Copy Number Change in the CNVR" or as a percentage of the total number of individuals characterized ($n = 1{,}190$), estimating the population frequency of structural variation in each genomic region.

We used simulation experiments to evaluate whether we had identified most, if not all, CNVRs detectable in our population with the Affymetrix GeneChip 100 and 500 K SNP genotyping platforms. Random sample sets of increasing size were selected from a pool of 1,190 subjects and CNVRs for each sample set were determined. Plotting the number of resultant CNVRs versus sample size revealed the effect of increasing sample size and showed our screen to be near saturation for this resolution of analysis (Fig. 2). The results follow a typical saturation curve defined by ($Y = \text{Max}*X/\text{Rate} + X$), where X and Y are points on the curve, Max is the maximum number of CNVRs in the population and Rate is the point at 50% saturation.

We tested the accuracy of CNAG copy number predictions in CNVRs #309 (loss), 336 (gain), 454 (gain) and 455 (loss). The estimated population frequencies of CNVs in these four genomic regions are 3.13, 2.02, 13.10 and 9.77%, respectively. For each CNVR, we tested 20 samples predicted by our computational approach to harbor a copy number alteration and 20 samples with diploid calls. The sensitivity and specificity of the CNAG copy number calls for CNVRs #309, #336, #454 and #455 were 90 and 85%, 80 and 100%, 60 and 80%, and 50 and 85%, respectively. These data suggest that the CNAG algorithm has greater

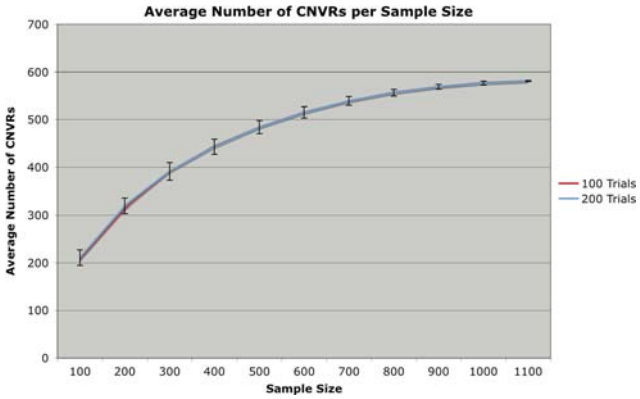

Fig. 2 Curves showing that we detected CNVRs in our population to saturation (given our SNP platform and computational detection and filtering criteria). Increasingly larger sample sets were randomly selected from a pool of 1,190 subjects and CNVRs were determined using the computational approaches described above. For each increment of 100 samples, the experiment was repeated either 100 (*red curve*) or 200 (*blue curve*) times and averages of the total number of CNVRs detected were calculated and plotted

detection sensitivity for copy number changes occurring at population frequencies <5%, which includes 99% of all CNVRs identified in our study. Consistent with previous reports (Huang et al. 2006), the specificity of the CNAG copy number calls was determined by quantitative PCR to be at least 80%, regardless of the population frequency of the copy number alteration. Together, these observations suggest that the CNV population frequencies reported in this study are accurate for CNVs occurring at frequencies of <5%, but may be underestimated for about 1% of CNVRs with genomic imbalances at population frequencies >5%.

Discussion

We used the Affymetrix GeneChip 100 and 500 K SNP genotyping platforms to assemble a genomic map of 578 CNV regions, covering approximately 220 Mb (7.3%) of the human genome. Although we observed CNVRs to be widespread across the human genome, copy number changes in the majority of these CNVRs are rare (>93% CNVRs include structural alterations occurring at <1% frequency). Population frequencies of 1–5% and >5% were estimated for CNVs present in approximately 6 and 1% of CNVRs, respectively. Our findings suggest that these structural variants are widespread across the human genome, but the majority occur at relatively low population frequencies. Therefore, based on the results obtained using the Affymetrix 100 and 500 K platforms to perform a genome-wide survey for CNVs, we postulate that most CNVs are infrequent events and are unlikely to seriously confound genome-wide case-control SNP association studies.

CNVs may be the product of recurrent events, resulting from non-allelic homologous recombination mediated by higher-order genomic architectural features (Freeman et al. 2006). The mapping of 34% of our CNVRs to genomic rearrangement hotspot sequences, namely sequences flanked by low copy repeats, supports this idea (Fig. 1). One limitation of the Affymetrix GeneChip 100 and 500 K SNP arrays is that these platforms are not tiling arrays, with a poor coverage over repetitive genomic regions where CNVs may be present. Since repetitive DNA stretches are often gene-poor, copy number alterations in these unmapped regions may represent benign genomic variants.

In keeping with evolutionary selection favoring genomic duplications over deletions (Freeman et al. 2006; Lee and Lupski 2006), regions with copy number gains were found to be more frequent in our population, more widespread across the genome, longer in size, and more likely to harbor genes. The frequency of copy number gains in the 578 CNVRs was 2.6-fold greater than the occurrence of deletions, with a total of 4,803 gains versus 1,867 deletions detected in 1,190 subjects. In addition, a larger fraction of the 578 unique CNVRs encompass gains (70%) compared to deletions (30%). Furthermore, CNVRs with deletions were found, on average, to be >2-fold shorter (207 vs. 494 kb for CNVRs encompassing DNA gains, *t*-test, $P < 0.0001$), while regions with copy number gains were more likely to include genes (Chi-square, 12.55, $P < 0.001$). In addition, an increased number of CNVRs with gains *vs.* deletions were associated with both OMIM morbid map (Fisher's Exact Test, two-tail $P < 0.001$) and known cancer genes (Fisher's Exact Test, two-tail $P < 0.05$) (Hamosh et al. 2002; Higgins et al. 2006; Sjoblom et al. 2006). We propose that increased copy numbers of tumor suppressor genes and hemizygosity for oncogenes may be protective against cancer.

Characterization of the CNVRs we identified revealed that these genomic regions contain hundreds of genes and disease loci. CNVRs are significantly enriched (hypergeometric test and Benjamini & Hochberg False Discovery Rate, $P < 0.05$) for rhodopsin-like receptor genes (e.g. olfactory, chemokine) (Supplementary Table 2). We found that 56% of the CNVRs are associated with known genes, including tumor suppressors and oncogenes. In fact, 55 known cancer genes overlap with CNVRs (Supplementary Table 3). Common CNVs rarely include dosage sensitive cancer genes, which may be the result of evolutionary selection. Of the 35 CNVRs associated with cancer genes, only three are associated with CNVs occurring at population frequencies >1%. These three CNVRs consist of copy number gains and overlap with the cyclin B1 interacting protein 1, Makorin-3, and myeloid/lymphoid or mixed-lineage leukemia genes. The presence of occasional structural variation involving cancer gene loci suggests a possible role for CNVs in genetic predisposition to cancer risk.

We found 172 genes in the OMIM morbid map of disease genes that overlap with the CNVRs we identified (Supplementary Table 4). Interestingly, one of these genes is *PMP22*, which causes Charcot-Marie-Tooth disease type 1A (CMT1A [MIM 118220]) by gene duplication (Lee and Lupski 2006). This observation supports evidence of previous reports describing cases of clinically healthy adult CMT1A patients (Berciano et al. 1994). Although the majority of CNVs mapping to disease loci occur at population frequencies <1%, we found 10 CNVRs with frequencies >1% associated with disease loci, including Prader-Willi syndrome (PWS [MIM 176270]), noninsulin-dependent diabetes mellitus (NIDDM [MIM 125853]), and autosomal recessive deafness (DFNB23 [MIM 609533]). We also detected rare copy number gains in a 401 kb genomic region (CNVR#284) that includes the *PRSS1* gene. Triplications of *PRSS1* (MIM 276000) have recently been associated with hereditary pancreatitis (Le Marechal et al. 2006), suggesting that the controls we identified with copy number gains in this locus may be at increased risk for this condition. CNVRs were also found to overlap with the *CCL3L1* (MIM 601395), *FCGR3B* (MIM 610665) and *HBD-2* (MIM 602215) genes, where increased germ-line genomic copy numbers have been recently associated with lower HIV, systemic lupus erythematosus (SLE), and Crohn's disease susceptibility, respectively (Gonzalez et al. 2005; Aitman et al. 2006; Fellermann et al. 2006). We estimated the frequency of copy number changes in CNVRs encompassing these three genes to be 3.3, 0.4 and 0.9% in our control population. In agreement with a recent report showing an average of 2.99 ± 1.74 (SD) copies of the *CCL3L1* gene in non-African subjects (Gonzalez et al. 2005), our predominantly Caucasian control group had an average copy number of 3 ± 0.6 (SD) in the genomic interval (CNVR #514) encompassing the *CCL3L1* locus. A *FCGR3B* locus copy number gain was identified in six individuals, while no genomic losses involving this region were observed, which is consistent with recent studies demonstrating a low *FCGR3B* copy number association with SLE and anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis (Aitman et al. 2006; Fanciulli et al. 2007). We observed a copy number gain of 4 ± 0.47 (SD) as the most frequent variation involving the *HBD-2* locus. Since the control subjects in our investigation do not have a known history of inflammatory bowel disease, our observation supports the recent report showing a relationship between predisposition to colonic Crohn's disease and a *HBD-2* gene copy of less than 4 (Fellermann et al. 2006).

We found that 68.6% of our CNVRs overlapped with recently reported CNVs (Iafrate et al. 2004; Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005; Conrad et al. 2006; Freeman et al. 2006; Hinds et al. 2006; McCarroll et al. 2006; Redon et al. 2006; Wong et al. 2007), providing

Table 3 Characteristics of the 18 CNVRs with frequently occurring (>2%) copy number changes

CNVR ID #	Position	Gain/ loss	Estimated population frequency	RefSeq genes	Observed in previous studies
454	chr14:19,272,965-20,000,915	GAIN	13.10%	OR4Q3; OR4M1; OR4N2; OR4K2; OR4K5; OR4K1; OR4K15; OR4K14; OR4K13; OR4L1; OR4K17; OR4N5; OR11G2; OR11H6; OR11H4; TTC5; CCNB1IP1; PARP2; TEP1; OSGEP; APEX1; TMEM55B	Sebat et al. (2004), Redon et al. (2006), Wong et al. (2007)
473	chr15:18,427,103-22,300,674	GAIN	11.12%	LOC283755; POTE15; OR4M2; OR4N4; LOC650137; TUBGCP5; CYFIP1; NIPA2; NIPA1; GOLGA8E; MKRN3; MAGEL2; NDN	Sebat et al. (2004), Sharp et al. (2005), Tuzun et al. (2005), Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
455	chr14:19,272,965-19,492,423	LOSS	9.77%	OR4Q3; OR4M1; OR4N2; OR4K2; OR4K5; OR4K1	Sebat et al. (2004), Redon et al. (2006), Wong et al. (2007)
517	chr17:40,758,788-42,166,282	GAIN	8.29%	ARHGAP27; LOC201175; PLEKHM1; C17orf69; CRHR1; IMP5; MAPT; STH; KIAA1267; LRRC37A; ARL17; LRRC37A2; NSF	Iafrate et al. (2004), Sebat et al. (2004), Sharp et al. (2005), Tuzun et al. (2005), Redon et al. (2006), Wong et al. (2007)
474	chr15:18,427,103-20,773,725	LOSS	8.01%	LOC283755; POTE15; OR4M2; OR4N4; LOC650137; TUBGCP5; CYFIP1; NIPA2; NIPA1	Sebat et al. (2004), Sharp et al. (2005), Tuzun et al. (2005), Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
86	chr2:242,634,423-242,730,382	LOSS	6.34%	–	Sharp et al. (2005), Tuzun et al. (2005), Redon et al. (2006)
348	chr10:44,530,696-47,485,249	GAIN	4.00%	RASSF4; C10orf10; C10orf25; ZNF22; OR13A1; ALOX5; MARCH8; ANUBL1; FAM21C; CTGLF1; PTPN20A; PTPN20B; PDZK5B; LOC594834; SYT15; GPRIN2; PPYR1; ANXA8L1; ANXA8	Iafrate et al. (2004), Sharp et al. (2005), Tuzun et al. (2005), Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
374	chr11:18,894,043-18,927,562	GAIN	3.87%	MRGPRX1	Tuzun et al. (2005), Redon et al. (2006), Wong et al. (2007)
456	chr14:21,120,353-22,511,019	GAIN	3.69%	OR10G2; OR4E2; DAD1; ABHD4; OXA1L; SLC7A7; MRPL52; MMP14; LRP10; REM2; RBM23; PRMT5; C14orf94; JUB	Sebat et al. (2004), Redon et al. (2006), Wong et al. (2007)
518	chr17:41,006,823-41,719,833	LOSS	3.55%	C17orf69; CRHR1; IMP5; MAPT; STH; KIAA1267	Sebat et al. (2004), Sharp et al. (2005), Redon et al. (2006), Wong et al. (2007)
514	chr17:31,310,858-33,248,878	GAIN	3.29%	CCL16; CCL14; CCL15; CCL23; CCL18; CCL3; CCL4; TBC1D3B; CCL3L3; CCL3L1; CCL4L1; CCL4L2; TBC1D3C; TBC1D3G; ZNHIT3; MYOHD1; PIGW; ZNF403; MGC4172; MRM1; LHX1; AATF; ACACA; C17orf78; TADA2L; DUSP14; AP1GBP1; DDX52; TCF2	Sharp et al. (2005), Redon et al. (2006), Wong et al. (2007)
309	chr8:137,757,137-137,955,330	LOSS	3.13%	–	Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
12	chr1:76,824,445-77,220,772	GAIN	2.82%	ST6GALNAC5	Redon et al. (2006), Wong et al. (2007)
328	chr9:42,050,602-44,108,554	GAIN	2.75%	ZNF658B	Sharp et al. (2005), Redon et al. (2006), Wong et al. (2007)
578	chr22:23,931,415-24,287,542	GAIN	2.64%	CRYBB2; LOC91353; LRP5L; ADRBK2	Sebat et al. (2004), Sharp et al. (2005), Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
295	chr8:12,285,367-12,286,403	LOSS	2.44%	–	Sebat et al. (2004), Sharp et al. (2005), Conrad et al. (2006), McCarroll et al. (2006), Redon et al. (2006), Wong et al. (2007)
496	chr16:32,411,529-35,003,380	GAIN	2.22%	TP53TG3; LOC649159	Iafrate et al. (2004), Sharp et al. (2005), Tuzun et al. (2005), Redon et al. (2006), Wong et al. (2007)
429	chr12:128,528,370-128,528,403	LOSS	2.18%	TMEM132D	Redon et al. (2006)

RefSeq genes mapping to each CNVR are shown. The estimated population frequency of copy number changes in each CNVR is indicated. Previous reports of CNVs in these genomic intervals are listed

further validation for our genomic map of CNVRs. Not surprisingly, there is a correlation ($r = 0.125$) between the population frequencies of copy number alterations within a CNVR and overlap with previously published CNVs, with more common CNVs being previously reported. In addition to validating the observations of recent smaller studies, our large population-based investigation permitted us to organize CNVs into CNVRs, to determine the population frequencies of genomic imbalances within the 578 CNVRs, and to identify 183 genomic regions (CNVRs) with novel CNVs (Supplementary Table 5). These novel findings include CNVRs #339 and #352, which harbor copy number alterations occurring at population frequencies of 2.0 and 1.7%, respectively. CNVR #339 includes copy number gains affecting the *orosomucoid 1* (MIM 138600) *and 2* (MIM 138610) genes, which encode for two acute phase plasma proteins with possible roles in immunosuppression, whereas CNVR #352 consists of genomic losses involving *protocadherin 15* (MIM 605514), a member of the *cadherin* gene superfamily. One possible mechanism by which CNVs may cause phenotypic diversity is by altering the expression of copy number variant genes. Although dosage effects may underlie the biological effects of certain CNVs, the functional consequences of copy number changes involving non-coding genomic sequences are not apparent. Further investigations are needed to distinguish biologically important structural variants from neutral polymorphisms.

Comparison of our CNVRs with those recently identified in 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap samples) (Redon et al. 2006) revealed 45% overlap (Fig. 1), suggesting that the two studies are complementary. In contrast to the former study, our investigation is sufficiently large to estimate the population frequencies of copy number changes and to identify less frequent variants. In addition, our genomic map is based on a homogeneous population, consisting of predominantly Caucasian (89.2%) controls from Ontario, Canada. Although, the 18 genomic regions where copy number changes occur most often (>2% frequency) were observed in both studies (Table 3), we identified hundreds of CNVRs, spanning at least 46 Mb (320 CNVRs), that were not detected in the HapMap samples (Supplementary Table 6). These include 9 CNVRs harboring genomic variants with population frequencies of 1–2%, which overlap with several genes, including genes belonging to the ras oncogene, cadherin, and acute phase plasma protein gene families. The remaining 311 CNVRs map to genomic regions where copy number changes are rare (<1% frequency).

In summary, using approximately 1,200 controls, we have demonstrated that CNVs are spread across hundreds of genomic regions covering 7.3% of the human genome. Although, we suggest that some degree of genome plasticity, manifested by CNVs, is expected in most individuals, copy number changes are infrequent events in the majority of CNVRs. We detected 183 novel CNVRs and suggest that certain CNVs may underlie genetic predisposition to cancer and other diseases. Using statistical simulation, we show that our sample size is sufficiently large to identify nearly all CNVRs present in our population using the Affymetrix GeneChip 100 and 500 K SNP genotyping platforms, and suggest that similar sample sizes will be necessary to fully characterize this relatively infrequent form of genetic variation in other worldwide populations. This North American population-based map of human copy number variable genomic regions will be a resource for future genetic studies.

Acknowledgments This work was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-96-011 (to SG & MC) and through cooperative agreements with members of the Colon Cancer Family Registry and P.I.s. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating institutions or investigators in the Colon CFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the Colon CFR. SG is the recipient of a grant from the Lustgarten Foundation for Pancreas Cancer Research, which also supported this work. Cancer Care Ontario, as the host organization to the ARCTIC Genome Project, acknowledges that this Project was funded by Genome Canada through the Ontario Genomics Institute, by Génome Québec, the Ministère du Développement Économique et Régional et de la Recherche du Québec and the Ontario Institute for Cancer Research. GZ is a Scholar of the Society of University Surgeons and a recipient of a Terry Fox Foundation Research Fellowship from the National Cancer Institute of Canada. The authors thank Dr. S. Ogawa for providing early access to CNAG version 2, Drs. C. Marshall and L. Feuk for their advice, Dr. D. Daftary, and Ms. T. Selander, Dr. Ling Liu and the Mount Sinai Hospital Biospecimen Repository for technical assistance.

References

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439:851–855

Berciano J, Calleja J, Combarros O (1994) Charcot-Marie-Tooth disease. Neurology 44:1985–1986

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81

Cotterchio M, McKeown-Eyssen G, Sutherland H, Buchan G, Aronson M, Easson AM, Macey J, Holowaty E, Gallinger S (2000) Ontario familial colon cancer registry: methods and first-year response rates. Chronic Dis Can 21:81–86

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, Shen MM, Kulp D, Kennedy GC, Mei R, Jones KW, Cawley S (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. Bioinformatics 21:1958–1963

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, Froguel P, Owen CJ, Pearce SH, Teixeira L, Guillevin L, Graham DS, Pusey CD, Cook HT, Vyse TJ, Aitman TJ (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721–723

Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn's disease of the colon. Am J Hum Genet 79:439–448

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 8:949–961

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA (2006) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55

Higgins ME, Claremount M, Major JE, Sander C, Lash AE (2006) Cancer Genes: a gene selection resource for cancer genome projects. Nucleic Acids Res 35:D721–D726

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85

Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA (2007) SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. Bioinformatics 23:57–63

Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, Shapero MH (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 7:83

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31:51–54

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:D493–D496

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996–1006

Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, Levy P, Ferec C (2006) Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 38:1372–1374

Lee JA, Lupski JR (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron 52:103–121

McCarroll S, Hadnott TH, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM, The International HapMap Consortium (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86–92

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65:6071–9

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H et al (2006) Global variation in copy number in the human genome. Nature 444:444–454

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D et al (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314:268–274

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732

Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104