BRIEF REPORT | Genetics inMedicine

Open

A high-resolution copy-number variation resource for clinical and population genetics

Mohammed Uddin, PhD[1,2], Bhooma Thiruvahindrapuram, MSc[1,2], Susan Walker, PhD[1,2], Zhuozhi Wang, PhD[1,2], Pingzhao Hu, PhD[1,2], Sylvia Lamoureux, BSc[1,2], John Wei, PhD[1,2], Jeffrey R. MacDonald, BSc[1,2], Giovanna Pellecchia, PhD[1,2], Chao Lu, PhD[1,2], Anath C. Lionel, PhD[1,2], Matthew J. Gazzellone, MSc[1,2], John R. McLaughlin, PhD[3–5], Catherine Brown, MSc[6], Irene L. Andrulis, PhD[5,7], Julia A. Knight, PhD[3,5], Jo-Anne Herbrick, BSc[1,2], Richard F. Wintle, PhD[1,2], Peter Ray, PhD[1,2,8], Dimitri J. Stavropoulos, PhD[8], Christian R. Marshall, PhD[1,2,8], and Stephen W. Scherer, PhD[1,2,9]

Purpose: Chromosomal microarray analysis to assess copy-number variation has become a first-tier genetic diagnostic test for individuals with unexplained neurodevelopmental disorders or multiple congenital anomalies. More than 100 cytogenetic laboratories worldwide use the new ultra-high resolution Affymetrix CytoScan-HD array to genotype hundreds of thousands of samples per year. Our aim was to develop a copy-number variation resource from a new population sample that would enable more accurate interpretation of clinical genetics data on this microarray platform and others.

Methods: Genotyping of 1,000 adult volunteers who are broadly representative of the Ontario population (as obtained from the Ontario Population Genomics Platform) was performed with the CytoScan-HD microarray system, which has 2.7 million probes. Four independent algorithms were applied to detect copy-number variations. Reproducibility and validation metrics were quantified using sample replicates and quantitative-polymerase chain reaction, respectively.

Results: DNA from 873 individuals passed quality control and we identified 71,178 copy-number variations (81 copy-number variations/individual); 9.8% (6,984) of these copy-number variations were previously unreported. After applying three layers of filtering criteria, from our highest confidence copy-number variation data set we obtained >95% reproducibility and >90% validation rates (73% of these copy-number variations overlapped at least one gene).

Conclusion: The genotype data and annotated copy-number variations for this largely Caucasian population will represent a valuable public resource enabling clinical genetics research and diagnostics.

Genet Med advance online publication 11 December 2014

Key Words: copy-number variation; CytoScan-HD; microarray; multiple congenital abnormalities; neurodevelopmental disorders

Copy-number variations (CNVs) constitute an abundant form of genetic variation and are increasingly being linked to genetic and phenotypic diversity as well as disease.[1–4] A wealth of literature exists for a significant role of CNVs in neurodevelopmental disorders and multiple congenital abnormalities.[5,6] For example, a review of 33 published studies by the International Standard Cytogenomic Array Consortium showed that ~12% of neurodevelopmental disorder cases can be explained by a CNV.[7] The clinical yield for autism spectrum disorders in recent studies shows that at least 5–15% of cases can be explained by CNVs that are either de novo or rarely inherited in nature.[8–10] Because most characterized penetrant CNVs are inherently rare, population-scale analyses are often required to assess relative disease risk and to elucidate the potential etiologic role of genetic events currently classified as "variants of unknown significance."[7]

The detection of CNVs in the clinical diagnostic setting is now largely based on an initial scan of the genome using microarrays to search for unbalanced alterations.[7,9,10] Locus-, gene-, and even exon-specific quantitative assays are also now used when a specific hypothesis is being pursued (e.g., when clinical assessment suggests a particular disease gene/mutation). In both instances, knowing the full spectrum of allelic architecture is necessary to make accurate clinical interpretations.[11] For these reasons, newer microarrays are being developed that contain dense probe content to allow robust testing for single-nucleotide polymorphism (SNP) genotypes and CNV detection. Dense SNP coverage allows zygosity testing, including assessment of uniparental disomy, as well as subpopulation structure analysis.

Recently, Affymetrix Corporation developed an array (CytoScan-HD) that consists of 2.7 million probes. Although

[1]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada; [2]Department of Genetics & Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; [3]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; [4]Public Health Ontario, Toronto, Ontario, Canada; [5]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada; [6]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada; [7]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada; [8]Division of Molecular Genetics, Department of Pediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada; [9]McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. Correspondence: Stephen W. Scherer (stephen.scherer@sickkids.ca)

Submitted 31 July 2014; accepted 7 November 2014; advance online publication 11 December 2014. doi:10.1038/gim.2014.178

these cover the entire genome, the densest representation is within genes; representation is even denser in known OMIM genes. In a recent study, high-resolution array assays in a small cohort of autism spectrum disorder and intellectual disability samples showed higher diagnostic yields and the capability to detect clinically relevant, smaller CNVs.[12] Moreover, in North America alone, more than 100 cytogenetic laboratories are now using the CytoScan-HD platform for both constitutional and cancer DNA testing. Recently, the CytoScan-Dx assay (the clinical name for the equivalent CytoScan-HD) obtained US Food and Drug Administration clearance for its use as a postnatal test for neurodevelopmental disorder or multiple congenital abnormality cases.

Having a large control series that is broadly representative of the underlying population that is genotyped with identical technology platforms provides the ideal situation for CNV calling.[13] Surprisingly, in the Database of Genomic Variants,[14] which is the standard resource used for CNV comparisons, only 44 population data sets from 55 studies are represented. Moreover, for these important studies, 41 different technology platforms have been used, and none of the data have yet been derived from the CytoScan-HD array. Here, we genotyped population-based samples from adult volunteers in the Ontario Population Genomics Platform (OPGP) using the CytoScan-HD array to generate the first such publicly available population data set. DNA and cell lines from this unique biological resource are also available for additional studies.

MATERIALS AND METHODS

The study was performed with direct participant consent and the approval of the research ethics boards at the Hospital for Sick Children and Mount Sinai Hospital, Toronto (studies 1000008876 and 06-0014-E, respectively).

OPGP sample collection

The OPGP consists of data and biospecimens collected from 2,690 adult volunteers from across Ontario, for whom recruitment was performed in two phases (see details for overall OPGP in **Supplementary Section S1** online **and Supplementary Tables S1–S3** online). Participants were first recruited through collaborations with the Ontario Familial Breast Cancer Registry (OFBCR)[15] and the Ontario Familial Colorectal Cancer Registry (OFCCR),[16] which are research resources used by international consortia in large studies of familial breast (OFBCR) and colorectal (OFCCR) cancers. These registries contain previously collected data and biospecimens from cancer patients, family members, and individuals from the general population who serve as controls. Population controls from these two registries were contacted and invited to participate in the OPGP and to reconsent so their previously collected data and biospecimens could be accessed by the OPGP. Reconsent was requested from a total of 1,886 controls from these registries, resulting in 1,462 controls being included in the OPGP (903 from the colorectal cancer registry and 559 from the breast cancer registry, for a 78% reconsent rate; see **Supplementary Table S1** online).

In the second phase, adult (age 20–79 years) volunteers residing across Ontario were recruited through a survey research process that involved random sampling from telephone directories, mailed introductions, and a combination of telephone interview and mailed questionnaires. Consenting individuals were mailed a package containing an explanatory letter, consent forms with a prepaid return envelope, and a blood kit for collection at a clinical laboratory in their community. The blood sample was sent via courier to the biospecimen repository at the Centre for Applied Genomics at the Hospital for Sick Children for transformation and DNA preparation (see details in **Supplementary Section S2** online). Of the 3,519 who completed the initial survey research process, blood sample collection kits were sent to all who consented (n = 2,074), among whom 1,228 (overall participation rate = 35%) returned both the specimen and signed consent form and were included in the OPGP.

DNA genotyping, CNV analysis, and quality control

DNA was genotyped (see **Supplementary Section S2** online) using the CytoScan-HD array following the manufacturer's protocol. The array consists of 2,696,550 probes that include 743,304 SNPs and 1,953,246 nonpolymorphic probes. The average probe spacing for RefSeq genes is 880 bp, and 96% of genes are represented. For this analysis, we have genotyped 1,000 samples; after extensive quality control (see **Supplementary Section S3** online), the OPGP subset for whom genotyping results are reported consists of 873 individuals and 22 sample replicates.

To achieve comprehensive CNV detection, we used four separate algorithms: Affymetrix Chromosome Analysis Suite (ChAS), iPattern, Nexus, and Partek. ChAS is the algorithm designed for use in clinical cytogenetic laboratories. Details of the other programs are found in **Supplementary Section S4** online. Our primary analysis was performed based on ChAS CNV calls, which were then supported using the remaining three algorithms to construct a set of high-confidence CNVs.[13] For all algorithms, we used eight probes and >1 kb as a minimum cutoff. Raw data from CNV genotyping are available in the NCBI database of Gene Expression Omnibus under accession no. GSE59150, and the CNV calls can be downloaded from http://www.tcag.ca/documents/projects/opgp873_chas.8p_1kb_one_replicates.txt.

Ancestry inference

To infer ancestry of the OPGP samples, we used 1,257 HapMap III samples (547,362 common SNPs with >95% call rates) as reference for 11 ethnically diverse populations[17] (see details in **Supplementary Section S5** online).

Experimental CNV validation and reproducibility

To examine the accuracy of our CNV calls, we used two different approaches. First, we used a CNV data set from 345 OPGP samples previously genotyped[18] using the lower-resolution Affymetrix genome-wide Human SNP 6.0 array. Second, we randomly selected 12 CNV regions in different size bins (ranging from 1.5 kb to 2.8 MB) and experimentally validated them using quantitative polymerase chain reaction (qPCR). Each qPCR assay was performed in triplicate for the test region and controls well-established to be diploid. The ratio of the average value for the test region to that for the control region had to be >1.4 or <0.7 for the CNV to be confirmed as a copy-number gain (duplication) or copy-number loss (deletion), respectively. In addition, the standard error of the ratio had to be <1.0 on the same scale for the assay to be considered reliable. To measure reproducibility of the microarray assay, we compared CNV calls from 22 randomly selected samples tested in replicate.

RESULTS

Ancestry determination

The majority of individuals in the OPGP cohort (95%) self-reported as being of European descent (**Figure 1a**), with the remaining 5% coming from African, Chinese, First Nations, Middle Eastern, South Asian, and South American backgrounds. Our inferred ancestry analysis using SNP genotypes shows strong concordance with the self-reported ancestry. The multidimensional scaling plot (**Figure 1b**) shows that the majority of the genotyped OPGP subset was highly clustered with the HapMap CEU population (Utah residents of Northern and Western European ancestry). The detected ancestry from PLINK analysis (**Figure 1c**) is also highly correlated: 94% Caucasian; 3.11% South American; 1.91% Asian; and 1% from an admixed population.

CNV distribution and reproducibility

After strict quality control, our final data set consisted of CNVs from 873 unrelated individuals (477 male and 396 female; mean age 58 years) (**Supplementary Figure S1** online; see demographics in **Supplementary Table S3** online). Because ChAS is the typical CNV detection program used for this array, we used it as our primary algorithm for CNV identification. Overall, we have not observed any frequency difference between males and females regarding common or rare CNV distribution (**Supplementary Table S4** online and **Supplementary Figure S2** online). As we have shown elsewhere,[13] to increase the sensitivity and specificity of CNV detection, we also used three other programs. CNV calls were stratified into three groups with increasingly stringent cutoffs: (i) "basic filter"—representing the entire CNV set exceeding at least 1 kb in length and having a minimum of eight consecutive probes; (ii) "research set"—a subset of the basic filter in which all the CNVs require the support of at least two algorithms (ChAS plus a second algorithm); and (iii) the "clinically stringent set," which includes CNVs with size and probe thresholds of 25 kb and 25 probes, respectively, for losses and of 50 kb and 50 probes, respectively, for gains (**Figure 2a**).

Applying the basic filter, we detected 71,178 CNVs, with the majority being losses (56,442) as compared with gains (14,736). CNV sizes ranged from 1 kb to 4.3 MB, with a median size of 9.95 kb (**Figure 2b** and **Supplementary Table S4** online).

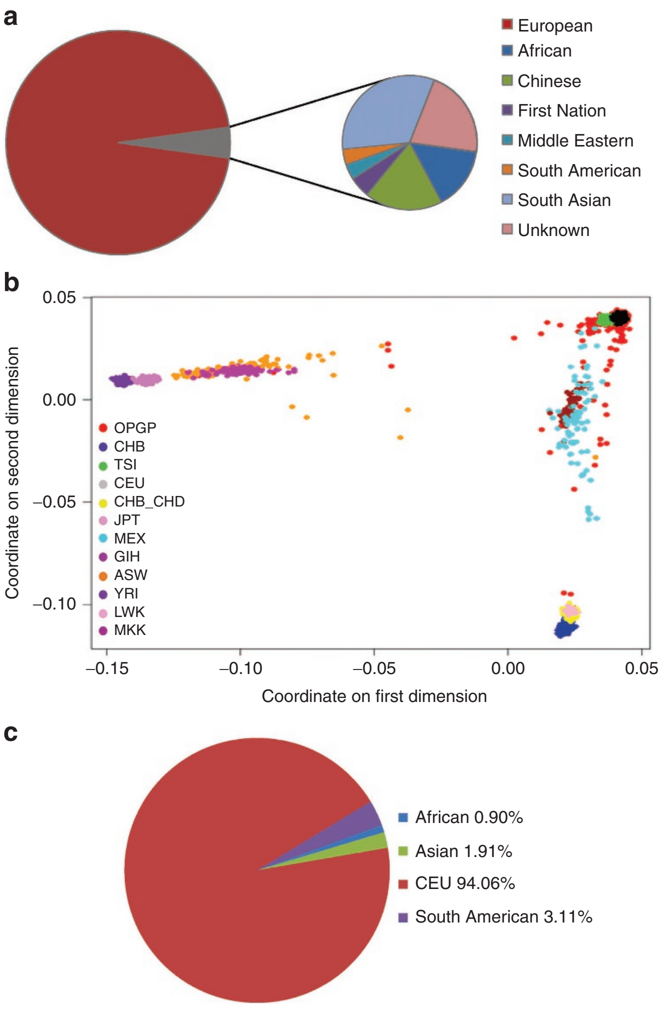

Figure 1 Ancestry determination among 873 adult volunteers in the Ontario Population Genomics Platform (OPGP). (**a**) Self-reported ethnic background of participants. (**b**) Multidimensional scaling (MDS) analysis of OPGP samples using HapMap III reference panel from 11 ethnically diverse populations. ASW, African ancestry in southwestern United States; CHB, Han Chinese individuals from Beijing, China; CEU, Utah residents of Northern and Western European ancestry; CHB_CHD, Chinese individuals from Beijing and Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, individuals from Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MKK, Maasai in Kinyawa, Kenya; MEX, individuals of Mexican ancestry in Los Angeles, California; TSI, Tuscans in Italy; YRI, Yoruba in Ibadan, Nigeria. The plot shows the distance between populations in a multidimensional scale using single-nucleotide polymorphisms as object. (**c**) Breakdown of the identified ethnic background for the OPGP cohort using identity by state (IBS) analysis. Each color represents a separate ethnic population.

Rare (<1% population frequency) large CNVs (>100 kb) comprised 5.2% of the CNVs detected. Male and female samples possessed 38,427 (54%) and 32,751 (46%) of CNVs, respectively. Importantly, 6,984 of the variants (mean size 7.7 kb) within the OPGP cohort are novel, having not been reported in any other studies (**Supplementary Table S5** online) within the Database of Genomic Variants.[14] This array is characterized by a high probe density for genic regions and, therefore, 62% of the detected CNVs overlapped with at least one gene.

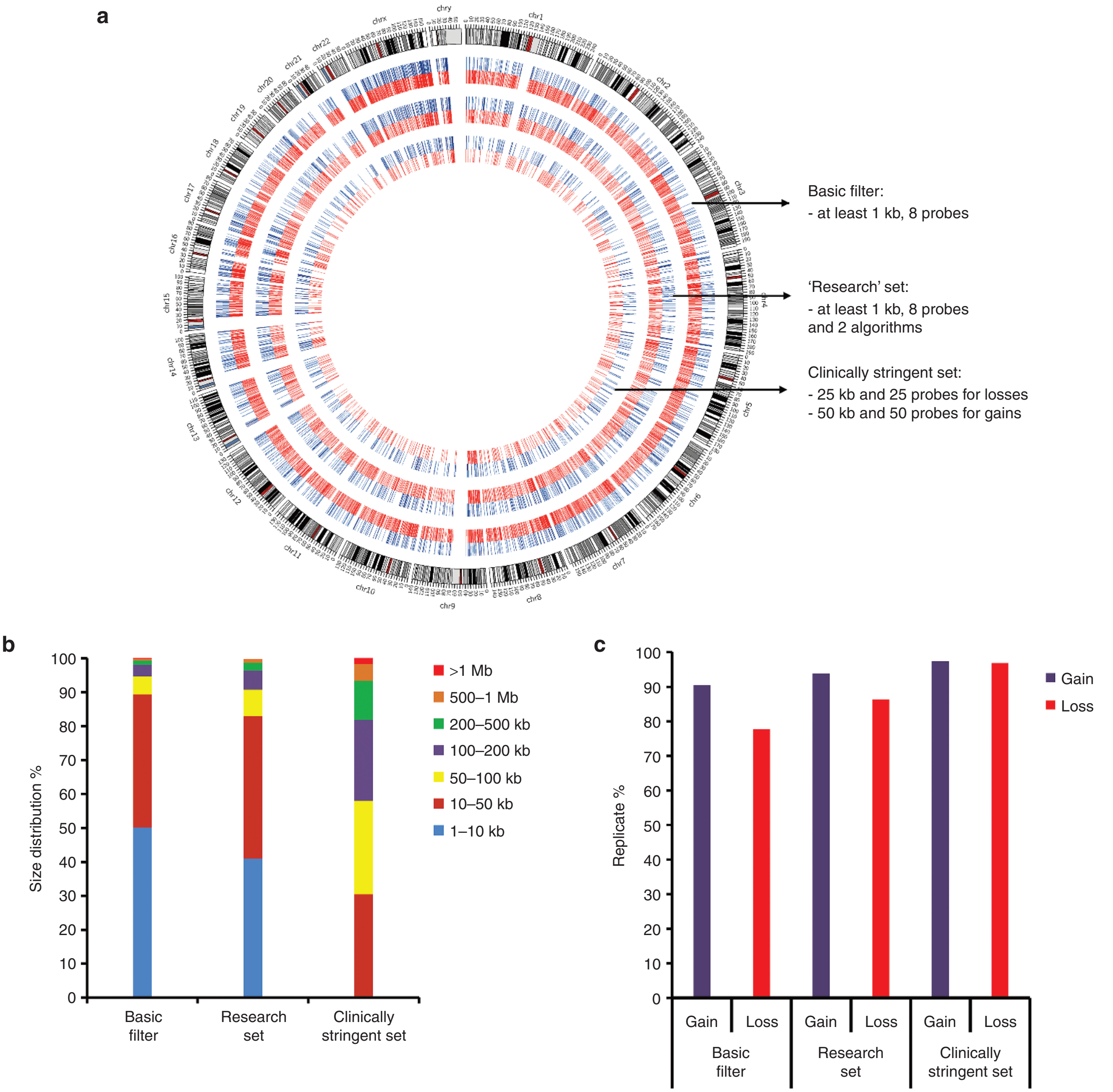

Figure 2 Copy-number variation (CNV) detection and characterization of 873 adult volunteers in the Ontario Population Genomics Platform (OPGP). (**a**) Genome-wide characterization of the detected CNV set. Chromosomes are shown outside of the circle and gains (blue) and losses (red) are shown in the inner rings. Each ring (in megabase (MB) scale) belongs to a classified group: "basic filter," "research set," or "clinically stringent set." (**b**) The size distribution of the CNVs (gains and losses) detected from the OPGP cohort. (**c**) The reproducibility of the CNVs analyzed from 22 replicates. The percentage of reproducibility in the plot is shown for gains (blue) and losses (red) for the three groups.

The reproducibility computed from 22 replicates (with at least 50% reciprocal overlaps) for the "basic filter" shows that >77% of CNVs (both losses and gains) are reproducible (**Figure 2c**).

After applying the "research set" filter, we obtained 34,502 CNVs (10,271 gains and 24,231 losses) with a median size of 13 kb (**Figure 2a**). The genic CNV rate remained unchanged (~62.7%), but the proportion of large (>100 kb) CNVs increased to 9.2% and reproducibility increased to 85% for both losses and gains.

By contrast, the "clinically stringent set" contained 6,965 high-confidence CNVs (2,576 gains and 4,389 losses) with a median size of 79 kb; 73% of CNVs within this specific tier are genic, and reproducibility is >96% for both losses and gains (**Figure 2a–c**). Comparison with the Affymetrix SNP array 6.0 data set showed that 81% of "research set" and 90% of "clinically stringent set" CNV calls were concordant between microarrays. Our qPCR validation set included 12 randomly chosen CNVs of different lengths from the “basic filter” CNV set, and 11 of 12 (91%) were validated by this method.

DISCUSSION

We present a new CNV resource derived from a North American population originating from Ontario, Canada. This is the first such public resource of data available for CNVs genotyped on the CytoScan-HD array. The resulting data should have tremendous value to guide diagnostic laboratories that are increasingly using the CytoScan-HD array or the Food and Drug Administration–approved CytoScan-Dx array to detect and assess the relevance of chromosomal abnormalities. In this study, we analyzed the CNV data in different stringency tiers (basic, research, and clinical filter) to facilitate investigation of research questions as well as for the appropriate clinical interpretation and prioritization of variants.

Our analysis found 6,984 CNVs not described previously. Many of these are small CNVs in the range of 1–15 kb that have previously been incompletely characterized (**Supplementary Table S6** online).[19] This higher-resolution analysis allows detection of novel CNVs affecting only small regions within a gene (e.g., *ADD2*; **Supplementary Figure S2** online) and helps to better define breakpoints of existing CNV calls (e.g. *PRIME2*; **Supplementary Figure S4** online).

Comparing (>70% reciprocal overlap) with the DECIPHER database, we found 10 OPGP samples harboring pathogenic gains and losses for five distinct genomic disorders (**Supplementary Table S7** online). For example, CNV genotyping of the OPGP samples detected variants overlapping the 16p13.11 region associated with male-biased neurodevelopmental disorders (**Supplementary Figure S5** online) as well as known disease-causing or risk genes (e.g., *PARK2*; **Supplementary Figure S6** online). In one example from our recent work,[20] isoform-specific small deletions within the *ASTN2/TRIM32* genes in males were implicated in neurodevelopmental disorders with diverse phenotypes. This segment of the genome is well represented on the CytoScan-HD array and, in fact, a smaller isoform-specific deletion was also detected in a male individual within the OPGP cohort.

Ultimately, high-resolution CNV calls using microarrays and sequencing will enable the construction of a chromosome imbalance map of the human genome. To best facilitate application in the clinical genetics setting, the data used for this map should be as accurate as possible and incorporate all geographic populations. In this work, we add valuable data and accompanying biospecimens to support such future clinical genetic research studies.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

The Ontario Population Genomics Platform was established with funding from the Centre for Applied Genomics (TCAG) and infrastructure support from the Canada Foundation for Innovation (CFI). Participant recruitment was made possible in collaboration with the Institute for Social Research at York University, the Ontario Familial Breast Cancer Registry (also supported by grant UM1 CA164920 from the US National Cancer Institute (NCI) and the Ontario Familial Colorectal Cancer Registry (also supported by grant UM1 CA167551 from the NCI and through a cooperative agreement U01/U24 CA074783)), for which we thank I.L.A., Michelle Cotterchio, Steve Gallinger, and Teresa Selander. The content of this article does not necessarily reflect the views or policies of the NCI, the host organizations, or collaborating centers, nor does mention of trade names, commercial products, or organizations imply endorsement by the government, funding agencies, host organizations, or collaborating centers. We thank Ting Wang, Kozue Otaka, and Guillermo Casallo for their help with the sample preparation. We also thank Alan H. Roter and Sam Dougaparsad from Affymetrix for their technical help. We thank the TCAG Science and Technology Innovation Centre, which is funded by Genome Canada and the Ontario Genomics Institute, the CFI, and the Ontario Research Fund of the Government of Ontario. The project was also supported by funds from the University of Toronto McLaughlin Centre and Genome Canada. S.W.S. holds the GlaxoSmithKline-CIHR Chair in Genome Sciences at the University of Toronto and the Hospital for Sick Children.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. Beckmann JS, Estivill X, Antonarakis SE. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* 2007;8:639–646.
2. Buchanan JA, Scherer SW. Contemplating effects of genomic structural variation. *Genet Med* 2008;10:639–647.
3. Conrad DF, Pinto D, Redon R, et al.; Wellcome Trust Case Control Consortium. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;464:704–712.
4. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 2010;12:e8.
5. Cook EH Jr, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. *Nature* 2008;455:919–923.
6. Pinto D, Pagnamenta AT, Klei L, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 2010;466:368–372.
7. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010;86:749–764.
8. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev* 2012;22:229–237.
9. Shen Y, Dies KA, Holm IA, et al.; Autism Consortium Clinical Genetics/DNA Diagnostics Collaboration. Clinical genetic testing for patients with autism spectrum disorders. *Pediatrics* 2010;125:e727–e735.
10. Stobbe G, Liu Y, Wu R, Hudgings LH, Thompson O, Hisama FM. Diagnostic yield of array comparative genomic hybridization in adults with autism spectrum disorders. *Genet Med* 2014;16:70–77.
11. Uddin M, Tammimies K, Pellecchia G, et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat Genet* 2014;46:742–747.
12. Qiao Y, Tyson C, Hrynchak M, et al. Clinical application of 2.7M Cytogenetics array for CNV detection in subjects with idiopathic autism and/or intellectual disability. *Clin Genet* 2013;83:145–154.
13. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* 2011;29:512–520.

14. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 2014;42 (database issue):D986–D992.

15. Figueiredo JC, Knight JA, Briollais L, Andrulis IL, Ozcelik H. Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the Breast Cancer Family Registry. *Cancer Epidemiol Biomarkers Prev* 2004;13:583–591.

16. Cotterchio M, McKeown-Eyssen G, Sutherland H, et al. Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis Can* 2000;21:81–86.

17. Altshuler DM, Gibbs RA, Peltonen L, et al.; International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* 2010;467:52–58.

18. Costain G, Lionel AC, Merico D, et al. Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays. *Hum Mol Genet* 2013;22:4485–4501.

19. Pang AW, Macdonald JR, Yuen RK, Hayes VM, Scherer SW. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. *G3 (Bethesda)* 2014;4:63–65.

20. Lionel AC, Tammimies K, Vaags AK, et al. Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Human Mol Genet* 2014;23: 2752–2768.