



Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts

John J. Connolly¹, Joseph T. Glessner^{1,2}, Berta Almoguera¹, David R. Crosslin³, Gail P. Jarvik³, Patrick M. Sleiman^{1,2} and Hakon Hakonarson^{1,2} *

¹ The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

² Department of Pediatrics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

³ Departments of Medicine (Medical Genetics) and Genome Sciences, University of Washington Medical Center, Seattle, WA, USA

Edited by:

Marylyn D. Ritchie, Pennsylvania State University, USA

Reviewed by:

Lifeng Tian, University of Pennsylvania, USA
Scott Matthew Williams, Dartmouth College, USA

*Correspondence:

Hakon Hakonarson, The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
e-mail: hakonarson@chop.edu

The goal of this paper is to review recent research on copy number variations (CNVs) and their association with complex and rare diseases. In the latter part of this paper, we focus on how large biorepositories such as the electronic medical record and genomics (eMERGE) consortium may be best leveraged to systematically mine for potentially pathogenic CNVs, and we end with a discussion of how such variants might be reported back for inclusion in electronic medical records as part of medical history.

Keywords: CNV, copy number, structural variation, eMERGE, review

WHAT ARE COPY NUMBER VARIATIONS?

Copy number variations (CNVs) are deletions and duplications in the genome that vary in length from ~50 base pairs to many megabases (50 base pair to 1 kilobase CNVs are typically considered indels). Events that cause CNVs include non-allelic homologous recombination, non-homologous end-joining, transposition of transposable elements, transposition of pseudogenes, variable numbers of tandem repeats, and replication errors following template-switching or fork stalling. CNVs are the primary mode by which an individual acquires a mutation, and occur at a rate of approximately 1.7×10^{-6} per locus as opposed to 1.8×10^{-8} for sequence variation (Lupski, 2007). Estimates of CNV frequency vary depending on the size of the structural variation classed as CNV – some estimates suggest that up to 12% of the genome may be variable in copy number, and that the cumulative result of CNV inheritance may constitute more than 10% of the human genome (Carter, 2007; Lupski et al., 2010). Recent studies suggest that the average human genome contains >1000 CNVs, covering approximately four million base pairs (Conrad et al., 2010; Mills et al., 2011), and occur at a rate of 0.07–0.12 per generation (Cordaux and Batzer, 2009; Itsara et al., 2010; Beck et al., 2011; Malhotra and Sebat, 2012). The Database of Genomic Variation (DGV)¹ currently lists over 100,000 published, unique, CNVs across the genome. While the majority continues to be benign, an increasing number of CNVs have been associated with disease susceptibility. Common functional consequences of CNVs typically demonstrate gene dose effect and include truncated protein sequences, eliminated/reduced protein expression (typically

the result of deletions), or increased protein expression (typically caused by duplications).

HOW ARE COPY NUMBER VARIATIONS IDENTIFIED? ARRAY-BASED APPROACHES

A range of approaches are available for detecting CNVs (Figure 1). The most common methods rely on computational methods, which leverage signals from genotyping and sequencing to infer CNVs. For example, large chromosomal anomalies can be detected through log R ratio (LRR) and B-allele frequency (BAF), data routinely generated and provided with single nucleotide polymorphism (SNP) and exome microarrays (e.g., Figure 2). For replication and validation, quantitative PCR – which compares the threshold cycles of a target versus reference sequence – is still widely deployed. In a similar vein, paralogs-ratio testing and molecular copy number counting are also used for validation.

For high-throughput CNV detection, the most common platforms are genome hybridization (CGH) arrays, genome-wide association (GWA) arrays, and second-generation sequencing (SGS). CGH arrays use artificial bacterial chromosomes or long synthetic oligonucleotides to probe either specific regions of interest or the entire genome (Greshock et al., 2007; Haraksingh et al., 2011). While this method has relatively low spatial resolution (typically >5–10 Mb; Kallioniemi et al., 1993) and requires a relatively large volume of DNA, CGH does offer high sensitivity and specificity (Greshock et al., 2007; Haraksingh et al., 2011), which is critical in a diagnostic context.

Single nucleotide polymorphism (SNP) arrays are more commonly used for CNV analysis, and CNVs can be identified from

¹<http://dgv.tcag.ca/dgv/app/home>

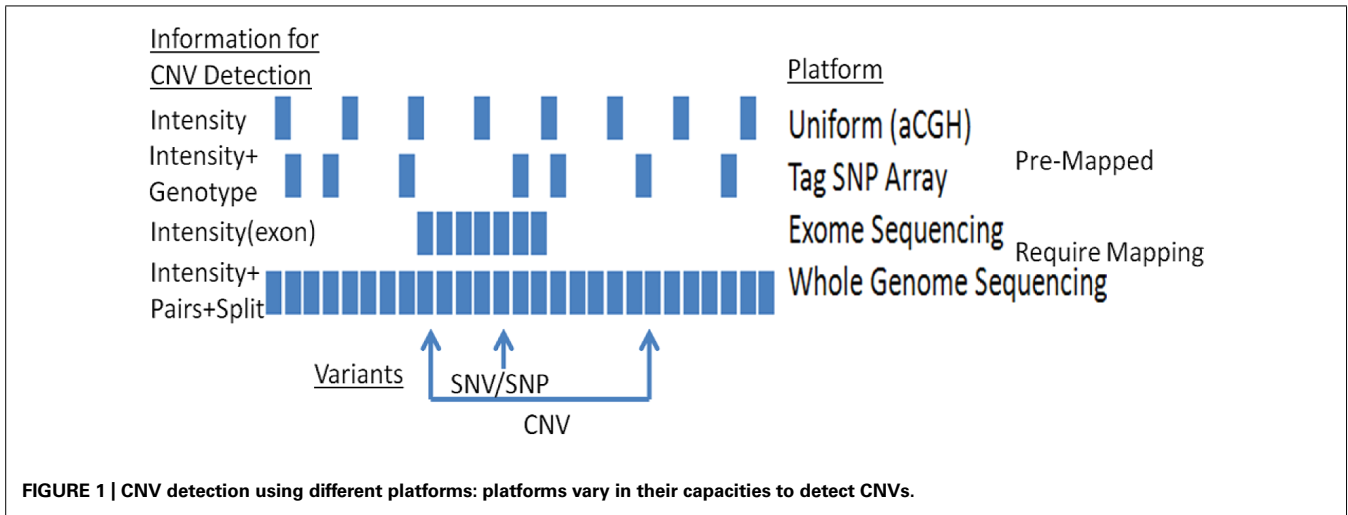


FIGURE 1 | CNV detection using different platforms: platforms vary in their capacities to detect CNVs.

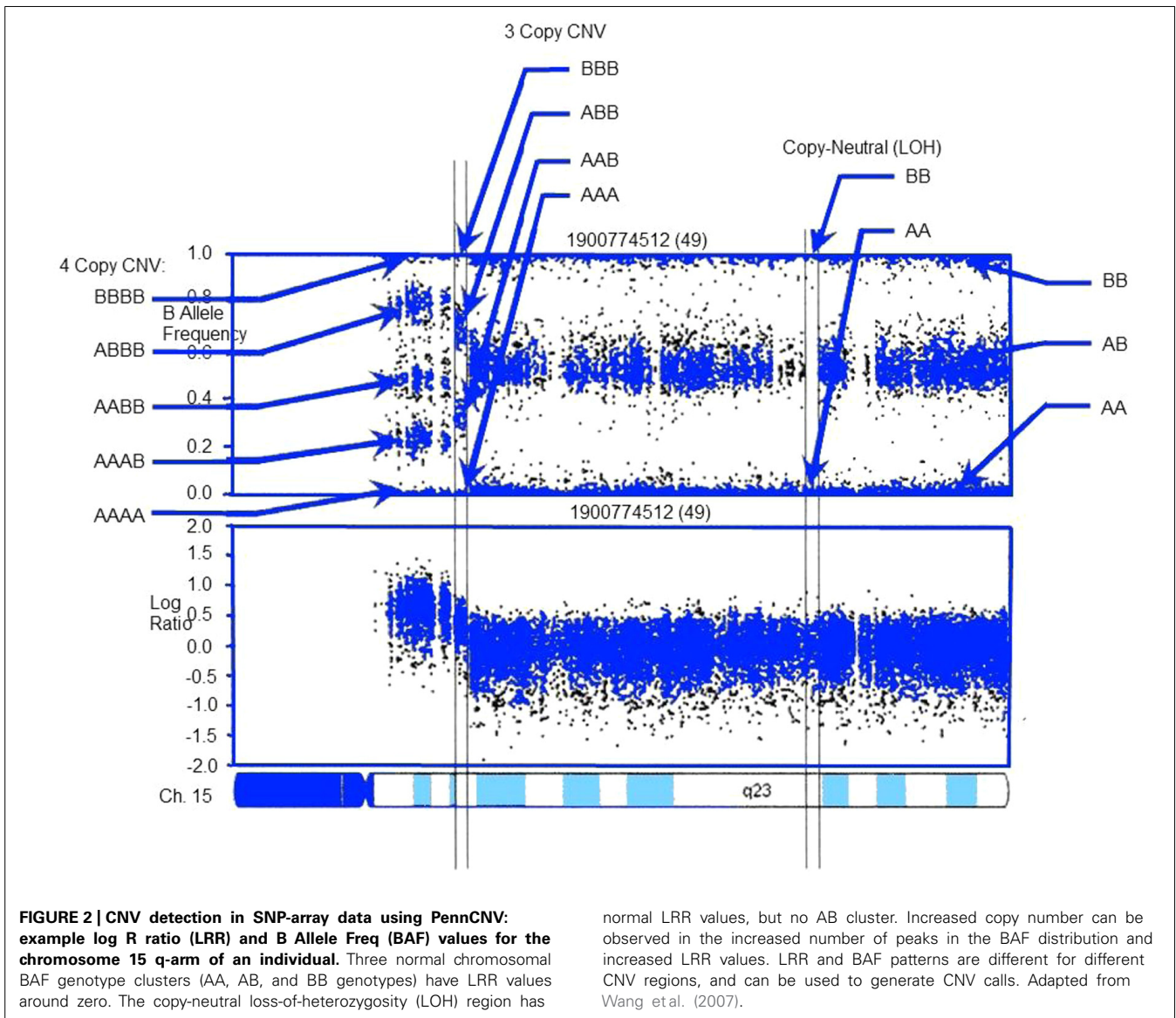


FIGURE 2 | CNV detection in SNP-array data using PennCNV: example log R ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual. Three normal chromosomal BAF genotype clusters (AA, AB, and BB genotypes) have LRR values around zero. The copy-neutral loss-of-heterozygosity (LOH) region has

normal LRR values, but no AB cluster. Increased copy number can be observed in the increased number of peaks in the BAF distribution and increased LRR values. LRR and BAF patterns are different for different CNV regions, and can be used to generate CNV calls. Adapted from Wang et al. (2007).

standard GWA array signals, or from arrays that utilize custom probes. Custom probes offer greater coverage of non-SNP sites, and can offer high sensitivity, particularly with regard to breakpoint resolution (Haraksingh et al., 2011). While conventional (i.e., non-custom) SNP arrays offer less specificity, they nevertheless represent a cost-effective option for characterizing CNVs and have been successfully applied to a wide range of phenotypes to date (Connolly and Hakonarson, 2012).

Importantly, it is possible to retroactively characterize CNVs from existing genome-wide association study (GWAS) data. In this context, the observed SNP signal of an allele relative to the normalized intensity of the allele can be used to deduce a deletion (decreased intensity) or duplication (increased intensity; Glessner et al., 2012). This possibility constitutes a major opportunity for custodians of large biorepositories such as electronic medical record and genomics (eMERGE), where a large volume of GWAS data has already been generated. Since its founding in 2007, the eMERGE consortium has produced dozens of GWASs on a range of phenotypes including lipids (Rasmussen-Torvik et al., 2012), arrhythmia (Ritchie et al., 2013), and white blood cell count (Crosslin et al., 2012) to name a few. For many of these phenotypes, no CNV studies have been published to date. This, we believe, represents an opportunity to identify new disease-associated loci without the generation of new genotype data, and will be addressed by the consortium in the immediate future. Similarly, we note that a large number of studies listed in the NHGRI GWAS catalog² do not have complementary CNV data, suggesting a largely under-utilized resource.

For array-based analyses, a range of packages are available. Both Affymetrix and Illumina – the two primary purveyors of SNP arrays – offer free software packages for CNV analysis. Independently developed toolsets are also available. These include circular binding segmentation (Olshen et al., 2004) MixHMM (Liu et al., 2010), GADA (Pique-Regi et al., 2008), PennCNV (Figure 2; Wang et al., 2007), and ParseCNV (Glessner et al., 2013a; the latter two were developed by eMERGE researchers and are widely used).

SEQUENCING-BASED APPROACHES

Common CNVs are well-covered by SNPs in existing arrays (Conrad et al., 2010; Wellcome Trust Consortium et al., 2010). However, a resequencing study by Pang et al. (2010) suggests that coverage of rare CNVs may be less comprehensive. The authors identified over 12,000 structural variants in 4,867 genes across 40 + mb of sequence (the Venter genome), which had been initially unreported. More than 24% of these CNVs would not have been imputed by SNP-association. Given that rare alleles can have large effect sizes and a high penetrance, these results underline the limitations of SNP arrays to identify certain pathogenic CNVs. SGS, which is far more proficient at identifying rare CNVs, offers an attractive solution in this regard – particularly in identifying novel insertions absent in the reference genome. This has obvious clinical utility. SGS also confers a number of other critical advantages in terms of ability to identify

smaller CNVs (<50 bp), and an enhanced capability for detecting breakpoints (Li and Olivier, 2013). Indeed, because SGS allows us to probe breakpoints at the level of base pairs, it facilitates capture of the signature of potential mutational mechanisms (Li and Olivier, 2013).

With SGS data, the most common methods for CNV identification from short-read analysis (Medvedev et al., 2010) are read-depth analysis (Xie and Tammi, 2009; Yoon et al., 2009; Abyzov et al., 2011), split-read mapping (Mills et al., 2006), paired-end read mapping (Korbel et al., 2009), and clone-based sequencing (Kidd et al., 2008). For all approaches, the most important determinants of accuracy are alignment and read-length. The average length of (reliable) reads is ~ from 100 to 150 bp, which is insufficient to eliminate erroneous mapping. As this metric improves, CNV-calling algorithms will become more accurate.

A large number of algorithms have been developed for indentifying CNVs from sequencing data, including CNVnator (Abyzov et al., 2011), PennCNV-Seq (in press), GenomeStrip (Handsaker et al., 2011), cnvHiTSeq (Bellos et al., 2012), and XHMM (Fromer et al., 2012). Different CNV algorithms have different strengths and weaknesses (see Li and Olivier, 2013 for review), and the most effective strategy in terms of minimizing erroneous CNV calls is to incorporate multiple toolsets, which can be validated computationally via local *de novo* assembly (e.g., see SVMerge, Wong et al., 2010).

DISEASE-ASSOCIATED COPY NUMBER VARIATIONS

As discussed elsewhere in this issue, GWASs have been successful in identifying common risk variants, particularly where the frequency of such variants is >5%. In addition to common variants, certain disorders have been shown to be enriched for rare CNVs (Conrad et al., 2010; Pang et al., 2010). In terms of functional impact, CNVs have been shown to be enriched in genes involved in immune responses, cell–cell signaling, and retrovirus- and transposition-related protein coding (Li and Olivier, 2013). A large number of phenotypes have now been associated with CNVs, including several rare diseases (Matsuura et al., 1997) and a range of neurodevelopmental disorders (Glessner et al., 2012), including depression (Glessner et al., 2010c), schizophrenia (Glessner et al., 2010b), and autism (Glessner et al., 2009). Autism provides a particularly good example of how our understanding of genetic risk factors and etiology is enhanced by CNV research, as demonstrated by a recent exome sequencing study (Iossifov et al., 2012) involving 343 families from the Simons Simplex Collection.

The study identified 59 “likely gene disruptions (LGDs)” in autism cases. Interestingly, the 59-strong LGD shared overlapped strongly with a set of 842 proteins that interact with the fragile X protein, FMRP. In total, 14 of the 59 LGDs encoded FMRP-interacting proteins ($P = 0.006$), as did 13 of 72 CNV candidates from the group’s previous CNV paper ($P = 0.0004$). Thus, 26 of 129 candidates were FMRP-related ($P < 1 \times 10^{-13}$). These results mark the fragile X mental retardation 1 (*FMRI*) gene as a high-profile autism candidate. Screening upstream targets of *FMRI*, the same group identified a deletion in *GRM5* that removes a single amino acid, causing an additional substitution at the same site. *GRM5* encodes the glutamate receptor mGluR5 (Bear et al., 2004),

²<http://www.genome.gov/gwastudies/>

which has been proposed as translational target in both ASD and ADHD (Elia et al., 2012; Silverman et al., 2012).

Several other CNV studies of autism have uncovered rare recurrent CNVs that have been informative. Our laboratory recently identified a range of CNVs in two major gene networks, ubiquitins and neuronal cell adhesion molecules that predispose to autism (Glessner et al., 2009). The ubiquitin–proteasome system is known to operate at pre- and post-synapses, and mediate neurotransmitter release, recycling of synaptic vesicles in pre-synaptic terminals, and modulating changes in dendritic spines and post-synaptic density (Yi and Ehlers, 2005). Neuronal cell adhesion molecules contribute to neurodevelopment by facilitating axon guidance, synapse formation and plasticity, and neuron–glial interactions.

Results from these and several other CNV studies suggest that genomic hotspots may be particularly vulnerable, which for autism include loci on chromosomes 1q21, 3p26, 15q11–q13, 16p11, and 22q11 (Bucan et al., 2009; Glessner et al., 2009; Pinto et al., 2010). Interestingly, these hotspots are part of large gene networks that are important to neural signaling and neurodevelopment, and have additionally been associated with other neuropsychiatric disorders. For example, studies of schizophrenia have highlighted structural mutations incorporating chromosomes 1q21, 15q13, and 22q11 (Glessner et al., 2010b). From an etiological perspective, autism and schizophrenia seem extremely different and it would seem counter-intuitive that associated loci should overlap. Some authors have addressed this peculiarity by proposing that the two disorders may in fact be opposite poles of the same spectrum (Crespi and Badcock, 2008). While such propositions await confirmation, they do highlight the potential of CNV studies to generate new hypotheses about the nature of complex diseases. Although individual structural variants explain relatively little by way of genetic variance, their cumulative is likely to be considerable. For autism, Marshall et al. (2008) suggested that CNVs play a causal role in 7% cases.

Beyond neuropsychiatric diseases, CNV studies have been published across a range of disease types, including heart disease (Goldmuntz et al., 2011), obesity (Glessner et al., 2010a), and cancer (Kuusisto et al., 2013). They have also recently been implicated in altered lifespan through alternative splicing mechanism (Glessner et al., 2013b).

COPY NUMBER VARIATIONS IN THE CONTEXT OF THE EMERGE CONSORTIUM

As illustrated in **Table 1**, the eMERGE consortium biorepository includes ~60,000 individuals that have been genotyped on high-density GWA arrays³, all of which have been linked with electronic medical records (EMRs). The size and diversity of the repository is such that it invokes the possibility for deep mining of disease-associated variants across multiple phenotypes. It is inevitable that a reasonable proportion of these individuals have disease-associated CNVs, and a larger proportion may be carriers of structural variants in recessive disease genes. By systematically characterizing CNVs across the biorepository, we have a very obvious opportunity to catalog CNVs and their disease-burden status.

We have now run PennCNV on eMERGE Phase I data (2007–2011), and will soon have circular binary segmentation analyses complete for the same set (50-kb to whole-chromosome). Relevant analyses will play a major role in the consortium's Phase II genomics program (2012–2015).

Similarly, the eMERGE consortium recently embarked upon a large-scale pharmacogenomics project [$n = \sim 9000$, review at Rasmussen-Torvik et al. (2012) in this issue], featuring a targeted sequencing platform developed by the Pharmacogenomics Research Network (PGRN), and covering 84 genes considered important for drug–gene interactions⁴. While the primary purpose of this project is to screen for existing pathogenic variants, this does offer an important opportunity to probe for novel variants in existing candidate genes, and to return results to patients' medical records. This clearly cannot be accomplished without paying heed to extensive medical, psychological, and ethical considerations, which are addressed elsewhere in this issue and in previous literature (Green et al., 2013). Assuming, however, that such considerations are adequately addressed, the section below considers how this might be accomplished and the potential to impact clinical care.

INTEGRATING CNVs WITH MEDICAL RECORDS – WHAT ARE THE OBSTACLES?

As discussed at length in this issue, the possibility of linking genomics data with EMRs represents a potentially major health-care opportunity. What variants/results and how to report them remains open to debate, and indeed part of the remit of the eMERGE consortium is to think through these hurdles.

An obvious first step is determining the pathogenicity of relevant CNVs. Traditionally (e.g., cytogenetics), interpretation of CNVs has concentrated on diseases where the mode of inheritance was dominant, and relied on simple case–control comparisons to discriminate pathogenic from non-pathogenic variations. Where the CNV was common (i.e., frequency >1–5%), it was typically classed as non-pathogenic. Thus, by process, “rare” implied “pathogenic.” With SGS and the increased capacity to detect smaller CNVs, this assumption falls down to a certain extent. We have started to see numerous studies where control and case *de novo* rate of small CNVs is as high as 5–10%. For rare CNVs in complex diseases, there is often insufficient power on which to base a judgment. Public databases that catalog pathogenic and non-pathogenic CNVs are therefore critical to determining frequencies of CNVs in disease cases and healthy controls.

Perhaps the most widely used catalog is the DGV, which aims to provide a “comprehensive summary of structural variation in the human genome” based on peer-review of relevant studies. While the DGV has obvious clinical and research relevance, several recent commentaries (Duclos et al., 2011; Hehir-Kwa et al., 2013) have urged caution in relying too heavily on its frequency and mapping statistics. As highlighted by Lee et al. (2007), many CNVs in the DGV are derived from single platforms/technologies, which may not necessarily translate to alternate approaches. Several recent studies (Perry et al., 2008; Conrad et al., 2010) suggest

³<http://www.genome.gov/27540473>

⁴www.pgrn.org

Table 1 | Summary of biorepositories and electronic medical records (EMRs) at 10 eMERGE-Institutions. Adapted from Gottesman et al. (2013).

| Institution | Biorepository | Recruitment model | Biorepository size | Race/ethnicity and age of donors |
|---------------------------------------|---|---------------------------------------|---|---|
| Boston Children's Hospital | Gene Partnership | Outpatient and hospital-based | 3,372 | 83% European 9% African 6% Asian 11% Hispanic/Latino Mean age: 23 years |
| Children's Hospital of Philadelphia | A Study of the Genetic Causes of Complex Pediatric Disorders | Population-based and disease-specific | 60,000 internal (plus 100,000 external) | 47.0% European 43.3% African 7.0% Admixed 1.7% Asian 0.8% Hispanic 0.2% Native Amer. Mean age: 11 years |
| Cincinnati Children's Hospital | Better Outcomes for Children | Outpatient and hospital-based | 8,472 | 73% European 10% African Mean age: 9 years |
| Geisinger Clinic | MyCode® | Population-based and disease-specific | 35,000 | 98% European Age: < 89 years |
| Group Health Seattle | ACT Study; Alzheimer's Disease Patient Registry (ADPR); Northwest Institute of Genetic Medicine (NWIGM) | Disease-specific and HMO-based | 5,859 | 92% European Age: > 50 years |
| Marshfield Clinic Research Foundation | Personalized Medicine Research Project | Population-based | 20,000 | 98% European Mean age: 48 years |
| Mayo Clinic | Vascular disease biorepository (VDB); Mayo Clinic Biobank; other disease-specific | Outpatient-based | 36,000 | 97% European Mean age: 63 years |
| Mount Sinai School of Medicine | BioMe™, The Charles Bronfman Institute for Personalized Medicine Biobank Program | Outpatient and hospital-based | 25,000 | 40% Hispanic/Latino 25% African 25% European |
| Northwestern University | NUgene | Outpatient and hospital-based | 12,000 | 9% Hispanic/Latino 12% African 78% European Mean age: 48 years |
| Vanderbilt University | BioVU | Outpatient and hospital-based | 155,000 | 2% Hispanic/Latino 15% African 80% European Mean age: 49 years |

that because of relatively low resolution in some studies, the size of relevant CNVs may be smaller than outlined in the DGV. Duclos et al. (2011) drew similar conclusions, stressing the “urgent need to validate the frequencies and boundaries of the CNVs recorded in the DGV.” This conclusion is based on the groups finding that some of the recorded CNVs are erroneously listed as polymorphic, which, if implemented in a medical setting may led to a deleterious CNV being called benign. Alternate CNV databases (e.g., dbVar; Lappalainen et al., 2013) have been established, but all are restrained by the quality of data on which they are based.

Other obstacles that have hampered development of CNV databases are inconsistent annotation of genomic data across studies, ill-defined curation protocols (e.g., QC-reporting, CNV-calling parameters), and incomplete phenotypic data. In each case, there is potential for consortium-led efforts to delineate best practices. To address the challenge of incomplete phenotypes, there is a particular opportunity for the eMERGE network. The majority of individuals enrolled in the eMERGE repository have their longitudinal EMRs linked to their genotype. This affords far greater potential for determining pathogenicity than traditional case-control studies, where controls may be

categorized as lacking a specific disease state, with no other phenotype data. Completeness-of-EMR is critical in this regard. For patients enrolled in the biorepository at The Children's Hospital of Philadelphia, the mean duration of EMRs is ~5.5 years, and is similar across other eMERGE sites. Relevant data include all ICD-9 diagnoses, lab values, procedures, and medications. Data of this length and depth should be considered minimal requirements for addressing pathogenicity on a large scale, while supplementation with disease-specific measures is also highly desirable.

Another major challenge in returning CNV data to patients' EMR concerns the nature of inheritance. An interesting study by Boone et al. (2013) recently sought to determine the rate of CNVs in recessive disease genes. The group used CGH to characterize deletion CNVs in 21,470 individual, identifying 3,212 heterozygous potential carrier deletions in 419 unique disease-associated genes. While many of these CNVs are likely benign polymorphisms, the group identified 206 heterozygous CNVs in multiple recessive genes, spanning 2–6 genes in each deletion. These CNVs, therefore, confer carrier status for multiple recessive conditions. Similarly, 307 individuals had multiple deletions in recessive disease genes. While many of these gene pairs

have unrelated function, a non-trivial proportion belongs to a shared pathway. Indeed, one participant had a CNV spanning three recessive immune genes *PSMB8*, *TAP1*, and *TAP2*, which are associated with autoinflammation, lipodystrophy, dermatosis syndrome (*PSMB8*), and type I bare lymphocyte syndrome (*TAP1* and *TAP2*). He also had a CNV in *CD19*, mutations of which are associated with common variable immunodeficiency. The authors were unable to determine whether the individual had a compromised immune system or presented with a history of immune disease (samples were anonymized). Nevertheless, he was clearly a multiple-deletion carrier, as were ~1.5% of the cohort: such information may be of direct clinical relevance to individuals' offspring – whether this should be shared remains open to debate.

Inherited CNVs pose a similar set of problems. While the majority of inherited CNVs may be in loci that lead to recessive disorders, this is not always the case. Indeed, one of the best-known CNVs is duplication at 15q11–q13, which accounts for up to 3% of autism cases (Sebat et al., 2007; Marshall et al., 2008). A complex scenario was recently described by Knijnenburg et al. (2009), where a child with a homozygous deletion in 15q13.3 (inherited from non-consanguineous, hemizygous carrier parents), resulted in hearing loss. Critically, if the CNV is a gain, three copies may have no phenotypic effect but four copies may have clinical consequences (Giorda et al., 2011). Conversely, when one parent carries a CNV loss in a recessive disease gene and the other parent carries a mutation in the same gene, this can result in compound heterozygosity in offspring (Hehir-Kwa et al., 2013; Paciorkowski et al., 2013). These findings stress the point that not only is the size, location, and direction of the CNV important, but so too is the number of copies. A range of other inheritance scenarios are reviewed by Hehir-Kwa et al. (2013), including X-linked CNVs (wide vary widely across individuals), and mosaic imbalances (Kousoulidou et al., 2013; may vary across an individual's cell types; Biesecker and Spinner, 2013; Forsberg et al., 2013).

Another point concerning CNV interpretation is the phenomenon of pleiotropy. As discussed above, a large proportion of reported recurrent CNVs have replicated *across* diseases (Cooper et al., 2011; Girirajan et al., 2011; Sahoo et al., 2011; Williams et al., 2011). Thus, the same microduplications at 1q21.1 have been associated with both autism and schizophrenia (Weiss et al., 2008; McCarthy et al., 2009). Relevant factors influencing the expressivity of this microduplication are a combination of environmental, epigenetic, and oligogenic (other modifier genes; Girirajan et al., 2010) factors. The precise mechanisms of causality that lead to a particular etiology are thus likely to be extremely complex, which calls into question what, if anything, might be reported in patients' EMRs. Such questions are the subject of ongoing debate (Fabsitz et al., 2010; Cassa et al., 2012), and are beyond the scope of this review. However, it is obvious that as genomic data becomes increasingly ubiquitous, we will require extensive guidelines in determining how CNV results should be interpreted and shared. For the same reason, it is critical that healthcare professionals receive adequate training and resources to understand and communicate test results.

Additionally, due to large numbers of cell divisions, CNVs, particularly deletions, can be acquired in the hematogenic progenitor cells. We have previously shown that acquired mosaicism increases with age and can be associated with hematological disorders (Laurie et al., 2012; Schick et al., 2013). However, when analyzing CNVs associated with neurological disorders, such acquired CNVs must be distinguished from germline mutations that are represented in non-hematological tissues, such as brain.

CONCLUSION

To date, a large number of diseases, across a large range of fields, have been associated with CNVs. We are still in our relative infancy in terms of deciding-upon the pathogenicity of such structural variants. We have stressed the need for a large, publicly accessible, and curated repository where CNVs that have been validated across platforms and technologies are stored. Whether this repository stems from improving existing catalogs or is developed *ab initio* remains to be determined, but the necessity of such a resource is compelling. Several eMERGE-led projects could funnel directly into such a repository, which would have real potential to impact healthcare.

A number of obstacles have stymied result-sharing – difficulties identifying CNVs (particularly in regions enriched for repetitive content), a shortage of standards, and the nature of CNV disease burden. These problems have attracted much attention in the past several years, and are well-characterized. While there is general agreement that such obstacles are substantial, there is a similar degree of optimism that benefits to be derived from solving these problems far outweigh the costs required. Again, consortium-led initiatives will likely be the most effective platforms for standardizing CNV-calling algorithms and developing guidelines for clinical care. The time is ripe for such initiatives, and we expect to see CNV-driven research make a major impact in clinical care in the next decade.

ACKNOWLEDGMENTS

This work was funded by institutional support from the Children's Hospital of Philadelphia and the National Human Genome Research Institute (# U01HG006830 and U01HG006375).

REFERENCES

- Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110
- Bear, M. E., Huber, K. M., and Warren, S. T. (2004). The mGluR theory of fragile X mental retardation. *Trends Neurosci.* 27, 370–377. doi: 10.1016/j.tins.2004.04.009
- Beck, C. R., Garcia-Perez, J. L., Badge, R. M., and Moran, J. V. (2011). LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.* 12, 187–215. doi: 10.1146/annurev-genom-082509-141802
- Bellos, E., Johnson, M. R., and Coin, L. J. M. (2012). cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.* 13, R120. doi: 10.1186/gb-2012-13-12-r120
- Biesecker, L. G., and Spinner, N. B. (2013). A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* 14, 307–320. doi: 10.1038/nrg3424
- Boone, P. M., Campbell, I. M., Baggett, B. C., Soens, Z. T., Rao, M. M., Hixson, P. M., et al. (2013). Deletions of recessive disease genes: CNV contribution to carrier states and disease-causing alleles. *Genome Res.* 23, 1383–1394. doi: 10.1101/gr.156075.113

- Bucan, M., Abrahams, B. S., Wang, K., Glessner, J. T., Herman, E. I., Sonnenblick, L. I., et al. (2009). Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.* 5:e1000536. doi: 10.1371/journal.pgen.1000536
- Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* 39, S16–S21. doi: 10.1038/ng2028
- Cassa, C. A., Savage, S. K., Taylor, P. L., Green, R. C., McGuire, A. L., Mandl, K. D., et al. (2012). Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res.* 22, 421–428. doi: 10.1101/gr.127845.111
- Connolly, J. J., and Hakonarson, H. (2012). The impact of genomics on pediatric research and medicine. *Pediatrics* 129, 1150–1160. doi: 10.1542/peds.2011-3636
- Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712. doi: 10.1038/nature08516
- Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., et al. (2011). A copy number variation morbidity map of developmental delay. *Nat. Genet.* 43, 838–846. doi: 10.1038/ng.909
- Cordaux, R., and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703. doi: 10.1038/nrg2640
- Crespi, B., and Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain. *Behav. Brain Sci.* 31, 241–261; discussion 261–320. doi: 10.1017/S0140525X08004214
- Crosslin, D. R., McDavid, A., Weston, N., Nelson, S. C., Zheng, X., Hart, E., et al. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE network. *Hum. Genet.* 131, 639–652. doi: 10.1007/s00439-011-1103-9
- Duclos, A., Charbonnier, F., Chambon, P., Latouche, J. B., Blavier, A., Redon, R., et al. (2011). Pitfalls in the use of DGV for CNV interpretation. *Am. J. Med. Genet. A* 155, 2593–2596. doi: 10.1002/ajmg.a.34195
- Elia, J., Glessner, J. T., Wang, K., Takahashi, N., Shtir, C. J., Hadley, D., et al. (2012). Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat. Genet.* 44, 78–84. doi: 10.1038/ng.1013
- Fabsitz, R. R., Fabsitz, R. R., McGuire, A., Sharp, R. R., Puggal, M., Beskow, L. M., et al. (2010). Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. *Circ. Cardiovasc. Genet.* 3, 574–580. doi: 10.1161/CIRCGENETICS.110.958827
- Forsberg, L. A., Absher, D., and Dumanski, J. P. (2013). Republished: non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. *Postgrad. Med. J.* 89, 417–426. doi: 10.1136/postgradmedj-2012-101322rep
- Fromer, M., Moran, J. L., Chambert, K., Banks, E., Bergen, S. E., Ruderfer, D. M., et al. (2012). Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* 91, 597–607. doi: 10.1016/j.ajhg.2012.08.005
- Giorda, R., Beri, S., Bonaglia, M. C., Spaccini, L., Scelsa, B., Manolagos, E., et al. (2011). Common structural features characterize interstitial intrachromosomal Xp and 18q triplications. *Am. J. Med. Genet. A* 155, 2681–2687. doi: 10.1002/ajmg.a.34248
- Girirajan, S., Brkanac, Z., Coe, B. P., Baker, C., Vives, L., Vu, T. H., et al. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.* 7:e1002334. doi: 10.1371/journal.pgen.1002334
- Girirajan, S., Rosenfeld, J. A., Cooper, G. M., Antonacci, F., Siswara, P., Itsara, A., et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* 42, 203–209. doi: 10.1038/ng.534
- Glessner, J. T., Bradfield, J. P., Wang, K., Takahashi, N., Zhang, H., Sleiman, P. M., et al. (2010a). A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am. J. Hum. Genet.* 87, 661–666. doi: 10.1016/j.ajhg.2010.09.014
- Glessner, J. T., Reilly, M. P., and Hakonarson, H. (2010b). Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10584–10859. doi: 10.1073/pnas.1000274107
- Glessner, J. T., Wang, K., Sleiman, P. M., Zhang, H., Kim, C. E., Flory, J. H., et al. (2010c). Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. *PLoS ONE* 5:e15463. doi: 10.1371/journal.pone.0015463
- Glessner, J. T., Connolly, J. J., and Hakonarson, H. (2012). Rare genomic deletions and duplications and their role in neurodevelopmental disorders. *Curr. Top. Behav. Neurosci.* 12, 345–360. doi: 10.1007/7854_2011_179
- Glessner, J. T., Li, J., and Hakonarson, H. (2013a). ParseCNV integrative copy number variation association software with quality tracking. *Nucleic Acids Res.* 41, e64. doi: 10.1093/nar/gks1346
- Glessner, J. T., Smith, A. V., Panossian, S., Kim, C. E., Takahashi, N., Thomas, K. A., et al. (2013b). Copy number variations in alternative splicing gene networks impact lifespan. *PLoS ONE* 8:e53846. doi: 10.1371/journal.pone.0053846
- Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., et al. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459, 569–573. doi: 10.1038/nature07953
- Goldmuntz, E., Paluru, P., Glessner, J., Hakonarson, H., Biegel, J. A., White, P. S., et al. (2011). Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. *Congenit. Heart Dis.* 6, 592–602. doi: 10.1111/j.1747-0803.2011.00582.x
- Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. *Genet. Med.* 15, 761–771. doi: 10.1038/gim.2013.72
- Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73
- Greshock, J., Feng, B., Nogueira, C., Ivanova, E., Perna, I., Nathanson, K., et al. (2007). A comparison of DNA copy number profiling platforms. *Cancer Res.* 67, 10173–10180. doi: 10.1158/0008-5472.CAN-07-2102
- Handsaker, R. E., Korn, J. M., Nemes, J., and McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat. Genet.* 43, 269–276. doi: 10.1038/ng.768
- Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E., and Snyder, M. (2011). Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PLoS ONE* 6:e27859. doi: 10.1371/journal.pone.0027859
- Hehir-Kwa, J., Pfundt, R., Veltman, J., and de Leeuw, N. (2013). Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clin. Genet.* 84, 415–421. doi: 10.1111/cge.12242
- Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299. doi: 10.1016/j.neuron.2012.04.009
- Itsara, A., Wu, H., Smith, J. D., Nickerson, D. A., Romieu, I., London, S. J., et al. (2010). De novo rates and selection of large copy number variation. *Genome Res.* 20, 1469–1481. doi: 10.1101/gr.107680.110
- Kallioniemi, O. P., Kallioniemi, A., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., et al. (1993). Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin. Cancer Biol.* 4, 41–46.
- Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. *Nature* 453, 56–64. doi: 10.1038/nature06862
- Knijnenburg, J., Oberstein, S. A., Frei, K., Lucas, T., Gijsbers, A. C., Ruivenkamp, C. A., et al. (2009). A homozygous deletion of a normal variation locus in a patient with hearing loss from non-consanguineous parents. *J. Med. Genet.* 46, 412–417. doi: 10.1136/jmg.2008.063685
- Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., et al. (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 10, R23. doi: 10.1186/gb-2009-10-2-r23
- Kousoulidou, L., Tanteles, G., Moutafi, M., Sismani, C., Patsalis, P. C., Anastasiadou, V., et al. (2013). 263.4 kb deletion within the TCF4 gene consistent with Pitt-Hopkins syndrome, inherited from a mosaic parent with normal phenotype. *Eur. J. Med. Genet.* 56, 314–318. doi: 10.1016/j.ejmg.2013.03.005
- Kuusisto, K. M., Akinrinade, O., and Schleutker, J. (2013). Copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS ONE* 8:e71802. doi: 10.1371/journal.pone.0071802

- Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., et al. (2013). DbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res.* 41, D936–D941. doi: 10.1093/nar/gks1213
- Laurie, C. C., Laurie, C. A., Rice, K., Doheny, K. F., Zelnick, L. R., McHugh, C. P., et al. (2012). Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* 44, 642–650. doi: 10.1038/ng.2271
- Lee, C., Iafrate, A. J., and Brothman, A. R. (2007). Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet.* 39, S48–S54. doi: 10.1038/ng2092
- Li, W., and Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiol. Genomics* 45, 1–16. doi: 10.1152/physiolgenomics.00082.2012
- Liu, Z., Li, A., Schulz, V., Chen, M., and Tuck, D. (2010). MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PLoS ONE* 5:e10909. doi: 10.1371/journal.pone.0010909
- Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. *Nat. Genet.* 39, S43–S47. doi: 10.1038/ng2084
- Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* 362, 1181–1191. doi: 10.1056/NEJMoa0908094
- Malhotra, D., and Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148, 1223–1241. doi: 10.1016/j.cell.2012.02.039
- Marshall, C. R., Noor, A., and Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* 82, 477–488. doi: 10.1016/j.ajhg.2007.12.009
- Matsuura, T., Sutcliffe, J. S., Fang, P., Galjaard, R. J., Jiang, Y. H., Benton, C. S., et al. (1997). De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat. Genet.* 15, 74–77. doi: 10.1038/ng0197-74
- McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., et al. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* 41, 1223–1227. doi: 10.1038/ng.474
- Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Res.* 20, 1613–1622. doi: 10.1101/gr.106344.110
- Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., et al. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16, 1182–1190. doi: 10.1101/gr.4565806
- Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65. doi: 10.1038/nature09708
- Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557–572. doi: 10.1093/biostatistics/kxh008
- Paciorkowski, A. R., Keppler-Noreuil, K., Robinson, L., Sullivan, C., Sajan, S., Christian, S. L., et al. (2013). Deletion 16p13.11 uncovers NDE1 mutations on the non-deleted homolog and extends the spectrum of severe microcephaly to include fetal brain disruption. *Am. J. Med. Genet. A* 161, 1523–1530. doi: 10.1002/ajmg.a.35969
- Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 11, R52. doi: 10.1186/gb-2010-11-5-r52
- Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., et al. (2008). The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.* 82, 685–695. doi: 10.1016/j.ajhg.2007.12.010
- Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466, 368–372. doi: 10.1038/nature09146
- Pique-Regi, R., Monso-Varona, J., Ortega, A., Seeger, R. C., Triche, T. J., Asgharzadeh, S., et al. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* 24, 309–318. doi: 10.1093/bioinformatics/btm601
- Rasmussen-Torvik, L. J., Pacheco, J. A., Wilke, R. A., Thompson, W. K., Ritchie, M. D., Kho, A. N., et al. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin. Transl. Sci.* 5, 394–399. doi: 10.1111/j.1752-8062.2012.00446.x
- Ritchie, M. D., Denny, J. C., Zuvich, R. L., Crawford, D. C., Schildcrout, J. S., Bastarache, L., et al. (2013). Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 127, 1377–1385. doi: 10.1161/CIRCULATIONAHA.112.000604
- Sahoo, T., Theisen, A., Rosenfeld, J. A., Lamb, A. N., Ravnani, J. B., Schultz, R. A., et al. (2011). Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. *Genet. Med.* 13, 868–880. doi: 10.1097/GIM.0b013e318217a06
- Schick, U. M., McDavid, A., Crane, P. K., Weston, N., Ehrlich, K., Newton, K. M., et al. (2013). Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS ONE* 8:e59823. doi: 10.1371/journal.pone.0059823
- Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., et al. (2007). Strong association of de novo copy number mutations with autism. *Science* 316, 445–449. doi: 10.1126/science.1138659
- Silverman, J. L., Smith, D. G., Rizzo, S. J., Karras, M. N., Turner, S. M., Tolu, S. S., et al. (2012). Negative allosteric modulation of the mGluR5 receptor reduces repetitive behaviors and rescues social deficits in mouse models of autism. *Sci. Transl. Med.* 4, 131ra51. doi: 10.1126/scitranslmed.3003501
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., et al. (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. doi: 10.1101/gr.6861907
- Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., et al. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* 358, 667–675. doi: 10.1056/NEJMoa075974
- Wellcome Trust Consortium, Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 464, 713–720. doi: 10.1038/nature08979
- Williams, N. M., Franke, B., Mick, E., Anney, R. J., Freitag, C. M., Gill, M., et al. (2011). Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am. J. Psychiatry* 169, 195–204.
- Wong, K., Keane, T. M., Stalker, J., and Adams, D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* 11:R128. doi: 10.1186/gb-2010-11-12-r128
- Xie, C., and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80. doi: 10.1186/1471-2105-10-80
- Yi, J. J., and Ehlers, M. D. (2005). Ubiquitin and protein turnover in synapse function. *Neuron* 47, 629–632. doi: 10.1016/j.neuron.2005.07.008
- Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 December 2013; accepted: 18 February 2014; published online: 18 March 2014.

Citation: Connolly JJ, Glessner JT, Almogueria B, Crosslin DR, Jarvik GP, Sleiman PM and Hakonarson H (2014) Copy number variation analysis in the context of electronic medical records and large-scale genomics consortium efforts. *Front. Genet.* 5:51. doi: 10.3389/fgene.2014.00051

This article was submitted to *Applied Genetic Epidemiology*, a section of the journal *Frontiers in Genetics*.

Copyright © 2014 Connolly, Glessner, Almogueria, Crosslin, Jarvik, Sleiman and Hakonarson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.