frontiers in GENETICS

Applied Genetic Epidemiology

Copy Number Variation Analysis in the Context of Electronic Medical Records & Large-Scale Genomics Consortium Efforts

John J Connolly, Joseph T Glessner, Berta Almoguera, David R Crosslin, Gail P Jarvik, Patrick Sleiman and Hakon Hakonarson

Journal Name:	Frontiers in Genetics
ISSN:	1664-8021
Article type:	Mini Review Article
Received on:	13 Dec 2013
Accepted on:	18 Feb 2014
Provisional PDF published on:	18 Feb 2014
Frontiers website link:	www.frontiersin.org
Citation:	Connolly JJ, Glessner JT, Almoguera B, Crosslin DR, Jarvik GP, Sleiman P and Hakonarson H(2014) Copy Number Variation Analysis in the Context of Electronic Medical Records & Large-Scale Genomics Consortium Efforts. *Front. Genet.* 5:51. doi:10.3389/fgene.2014.00051
Article URL:	/Journal/Abstract.aspx?s=1265&name=applied%20genetic%20epidemiology&ART_DOI=10.3389/fgene.2014.00051 (If clicking on the link doesn't work, try copying and pasting it into your browser.)
Copyright statement:	© 2014 Connolly, Glessner, Almoguera, 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.

Title: Copy Number Variation Analysis in the Context of Electronic Medical Records & Large-Scale Genomics Consortium Efforts

Authors: John J Connolly[1], Joseph T Glessner[1,2], Berta Almoguera[1], David R Crosslin[3], Gail P Jarvik[3], Patrick Sleiman[1,2] & Hakon Hakonarson[1,2]

Affiliations:

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

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

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

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

Abstract:

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.

1. What are CNVs?

CNVs are deletions and duplications in the genome that vary in length from ~~>1k~~~50 base pairs to many megabases. 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[1]. 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[2,3]. Recent studies suggest that the average human genome contains >1000 CNVs, covering approximately 4 million base pairs[4,5], and occur at a rate of .07-

0.12 per generation[6-9]. The Database of Genomic Variation (DGV, http://dgv.tcag.ca/dgv/app/home) 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).

2. How are CNVs 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 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. CGH arrays use artificial bacterial chromosomes or long synthetic oligonucleotides to probe either specific regions of interest or the entire genome[10,11]. While this method has relatively low spatial resolution (typically >5–10 Mb)[12] and requires a relatively large volume of DNA, CGH does offer high sensitivity and specificity[10,11], 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 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[11]. 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[13].

Importantly, it is possible to retroactively characterize CNVs from existing 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)[14]. This possibility constitutes a major opportunity for custodians of large biorepositories such as 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[15], arrhythmia[16], and white blood cell count[17] 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 catalog (http://www.genome.gov/gwastudies/) 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[18] MixHMM[19], GADA[20], PennCNV (**Figure 2**)[21], and ParseCNV[22] (the latter two were developed by eMERGE researchers and are widely used).

INSERT FIGURE 1 ~HERE

Sequencing-Based Approaches: Common CNVs are well-covered by SNPs in existing arrays (Wellcome Trust Consortium (2010)[23], Conrad et al. (2010)[5]). However, a resequencing study by Pang et al. (2010)[24] 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. Second generation sequencing (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 (<50bp), and an enhanced capability for detecting breakpoints[25]. Indeed, because SGS allows us to probe breakpoints at the level of base pairs, it facilitates capture of the signature of potential mutational mechanisms[25].

With SGS data, the most common methods for CNV identification from short-read analysis[26] are read-depth analysis[27-29], split-read mapping[30], paired-end read mapping[31], and clone-based sequencing[32]. For all approaches, the most important determinants of accuracy are alignment and read-length. The average length of (reliable) reads is ~100-150bp, 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[27], PennCNV-Seq (in press), GenomeStrip[33], cnvHiTSeq[34], and XHMM[35]. Different CNV algorithms have different strengths and weaknesses (see Li & Olivier[25] 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 [PMID:21194472)].

INSERT FIGURE 2 ~HERE

3. Disease-Associated CNVs:

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[5,24]. 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[25]. A large number of phenotypes have now been associated with CNVs, including several rare diseases[36] and a range of neurodevelopmental disorders[14], including depression[37], schizophrenia[38], and autism[39]. 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[40] involving 343 families from the Simons Simplex Collection.

Comment [JC1]: Will link to papers from Iftikhar … (pending submission)

The study identified 59 "likely gene disruptions (LGD)" 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<1x10^{-13}). These results mark the fragile X mental retardation 1 (*FMR1*) gene as a high-profile autism candidate. Screening upstream targets of *FMR1*, 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[41], which has been proposed as translational target in both ASD and ADHD[42,43].

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[39]. 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[44]. 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[39,45,46]. 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[38]. 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[47]. 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)[48] 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[49], obesity[50], and cancer[51]. They have also recently been implicated in altered lifespan through alternative splicing mecahnism[52].

4. CNVs 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 arrays (review at http://www.genome.gov/27540473), 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).

INSERT TABLE 1 ~HERE

Similarly, the eMERGE consortium recently embarked upon a large-scale pharmacogenomics project (n = ~9000, review at Rasmussen-Torvik et al. in this issue), featuring a targeted sequencing platform developed by the Pharmacogenomics Research Network (PGRN), and covering 84 genes considered important for drug-gene interactions (www.pgrn.org). 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[53]. Assuming, however, that such considerations are adequately addressed, the section below considers how this might be accomplished and the potential to impact clinical care.

Comment [JC2]: Will link to Kullo et al. (pending submission)

5. 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 healthcare 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.

Comment [JC3]: Again, will link to Kullo et al. (pending submission)

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[54,55] have urged caution in relying too heavily on its frequency and mapping statistics. As highlighted by Lee et al. (2007)[56], many CNVs in the DGV are derived from single platforms/technologies, which may not necessarily translate to alternate approaches. Several recent studies[5,57] suggest 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)[54] 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[58]) 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 ~~a~~ 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[48,59]. A complex scenario was recently described by Knijnenburg et al. (2009)[60], 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[61]. 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[55,62]. These findings stress the point that not only is the size, location, and direction of the CNV important~~ce~~, but so too is the number of copies. A range of other inheritance scenarios are reviewed by Hehir-Kwa et al. (2013)[55], including X-linked CNVs (wide vary widely across individuals), and mosaic imbalances[63] (may vary across an individual's cell types[64,65]).

Another point concerning CNV interpretation is the phenomenon of pleiotropy. As discussed above, a large proportion of reported recurrent CNVs have replicated *across* diseases[66-69]. Thus, the same microduplications at 1q21.1 have been associated with both autism and schizophrenia[70,71]. Relevant factors influencing the expressivity of this microduplication are a combination of environmental, epigenetic, and oligogenic (other modifier genes)[72] 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[73,74], 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[75,76]. 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.

Conclusions:

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.

1. Lupski, J.R. Genomic rearrangements and sporadic disease. *Nat Genet* **39**, S43-7 (2007).
2. Lupski, J.R. et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362**, 1181-91 (2010).
3. Carter, N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet* **39**, S16-21 (2007).
4. Mills, R.E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).
5. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-12 (2010).
6. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-41 (2012).
7. Itsara, A. et al. De novo rates and selection of large copy number variation. *Genome Res* **20**, 1469-81 (2010).
8. Cordaux, R. & Batzer, M.A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691-703 (2009).
9. Beck, C.R., Garcia-Perez, J.L., Badge, R.M. & Moran, J.V. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet* **12**, 187-215 (2011).
10. Greshock, J. et al. A comparison of DNA copy number profiling platforms. *Cancer Res* **67**, 10173-80 (2007).
11. Haraksingh, R.R., Abyzov, A., Gerstein, M., Urban, A.E. & Snyder, M. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PLoS One* **6**, e27859 (2011).
12. Kallioniemi, O.P. et al. Comparative genomic hybridization: a rapid new method for detecting and mapping DNA amplification in tumors. *Semin Cancer Biol* **4**, 41-6 (1993).
13. Connolly, J.J. & Hakonarson, H. The impact of genomics on pediatric research and medicine. *Pediatrics* **129**, 1150-60 (2012).
14. Glessner, J.T., Connolly, J.J. & Hakonarson, H. Rare Genomic Deletions and Duplications and their Role in Neurodevelopmental Disorders. *Curr Top Behav Neurosci* (2012).
15. Rasmussen-Torvik, L.J. et al. High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin Transl Sci* **5**, 394-9 (2012).
16. Ritchie, M.D. et al. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* **127**, 1377-85 (2013).
17. Crosslin, D.R. et al. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum Genet* **131**, 639-52 (2012).
18. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-72 (2004).
19. Liu, Z., Li, A., Schulz, V., Chen, M. & Tuck, D. MixHMM: inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. *PLoS One* **5**, e10909 (2010).
20. Pique-Regi, R. et al. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309-18 (2008).
21. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).
22. Glessner, J.T., Li, J. & Hakonarson, H. ParseCNV integrative copy number variation association software with quality tracking. *Nucleic Acids Res* **41**, e64 (2013).

23. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20 (2010).
24. Pang, A.W. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol* **11**, R52 (2010).
25. Li, W. & Olivier, M. Current analysis platforms and methods for detecting copy number variation. *Physiol Genomics* **45**, 1-16 (2013).
26. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res* **20**, 1613-22 (2010).
27. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**, 974-84 (2011).
28. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
29. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res* **19**, 1586-92 (2009).
30. Mills, R.E. et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**, 1182-90 (2006).
31. Korbel, J.O. et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10**, R23 (2009).
32. Kidd, J.M. et al. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56-64 (2008).
33. Handsaker, R.E., Korn, J.M., Nemesh, J. & McCarroll, S.A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269-76 (2011).
34. Bellos, E., Johnson, M.R. & LJ, M.C. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol* **13**, R120 (2012).
35. Fromer, M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* **91**, 597-607 (2012).
36. Matsuura, T. et al. De novo truncating mutations in E6-AP ubiquitin-protein ligase gene (UBE3A) in Angelman syndrome. *Nat Genet* **15**, 74-7 (1997).
37. Glessner, J.T. et al. Duplication of the SLIT3 locus on 5q35.1 predisposes to major depressive disorder. *PLoS One* **5**, e15463 (2010).
38. Glessner, J.T. et al. Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc Natl Acad Sci U S A* **107**, 10584-9 (2010).
39. Glessner, J.T. et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569-73 (2009).
40. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
41. Bear, M.F., Huber, K.M. & Warren, S.T. The mGluR theory of fragile X mental retardation. *Trends Neurosci* **27**, 370-7 (2004).
42. Silverman, J.L. et al. Negative allosteric modulation of the mGluR5 receptor reduces repetitive behaviors and rescues social deficits in mouse models of autism. *Sci Transl Med* **4**, 131ra51 (2012).
43. Elia, J. et al. Genome-wide copy number variation study associates metabotropic glutamate receptor gene networks with attention deficit hyperactivity disorder. *Nat Genet* **44**, 78-84 (2012).

44. Yi, J.J. & Ehlers, M.D. Ubiquitin and protein turnover in synapse function. *Neuron* **47**, 629-32 (2005).
45. Bucan, M. et al. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet* **5**, e1000536 (2009).
46. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
47. Crespi, B. & Badcock, C. Psychosis and autism as diametrical disorders of the social brain. *Behav Brain Sci* **31**, 241-61; discussion 261-320 (2008).
48. Marshall, C.R. et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88 (2008).
49. Goldmuntz, E. et al. Microdeletions and microduplications in patients with congenital heart disease and multiple congenital anomalies. *Congenit Heart Dis* **6**, 592-602 (2011).
50. Glessner, J.T. et al. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am J Hum Genet* **87**, 661-6 (2010).
51. Kuusisto, K.M. et al. Copy Number Variation Analysis in Familial BRCA1/2-Negative Finnish Breast and Ovarian Cancer. *PLoS One* **8**, e71802 (2013).
52. Glessner, J.T. et al. Copy number variations in alternative splicing gene networks impact lifespan. *PLoS One* **8**, e53846 (2013).
53. Green, R.C. et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* (2013).
54. Duclos, A. et al. Pitfalls in the use of DGV for CNV interpretation. *Am J Med Genet A* **155A**, 2593-6 (2011).
55. Hehir-Kwa, J., Pfundt, R., Veltman, J. & de Leeuw, N. Pathogenic or not? Assessing the clinical relevance of copy number variants. *Clin Genet* (2013).
56. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* **39**, S48-54 (2007).
57. Perry, G.H. et al. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82**, 685-95 (2008).
58. Lappalainen, I. et al. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* **41**, D936-41 (2013).
59. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9 (2007).
60. Knijnenburg, J. et al. A homozygous deletion of a normal variation locus in a patient with hearing loss from non-consanguineous parents. *J Med Genet* **46**, 412-7 (2009).
61. Giorda, R. et al. Common structural features characterize interstitial intrachromosomal Xp and 18q triplications. *Am J Med Genet A* **155A**, 2681-7 (2011).
62. Paciorkowski, A.R. et al. 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* **161A**, 1523-30 (2013).
63. Kousoulidou, L. et al. 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**, 314-8 (2013).
64. Forsberg, L.A., Absher, D. & Dumanski, J.P. Republished: Non-heritable genetics of human disease: spotlight on post-zygotic genetic variation acquired during lifetime. *Postgrad Med J* **89**, 417-26 (2013).
65. Biesecker, L.G. & Spinner, N.B. A genomic view of mosaicism and human disease. *Nat Rev Genet* **14**, 307-20 (2013).

66. Cooper, G.M. et al. A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838-46 (2011).
67. Girirajan, S. et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet* **7**, e1002334 (2011).
68. Sahoo, T. et al. Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. *Genet Med* **13**, 868-80 (2011).
69. Williams, N.M. et al. 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* (2011).
70. McCarthy, S.E. et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-7 (2009).
71. Weiss, L.A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667-75 (2008).
72. Girirajan, S. et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42**, 203-9 (2010).
73. Cassa, C.A. et al. Disclosing pathogenic genetic variants to research participants: quantifying an emerging ethical responsibility. *Genome Res* **22**, 421-8 (2012).
74. Fabsitz, R.R. et al. 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**, 574-80 (2010).
75. Laurie, C.C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* **44**, 642-50 (2012).
76. Schick, U.M. et al. Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS One* **8**, e59823 (2013).
77. Gottesman, O. et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* (2013).

Table 1, Summary of biorepositories and electronic medical records (EMRs) at 10 eMERGE-Institutions. Adapted from Gottesman et al. (2013)[77].

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
GroupHealth 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	Bio*Me*™, 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

Figure 1, CNV detection using different platforms: Platforms vary in their capacities to detect CNVs.

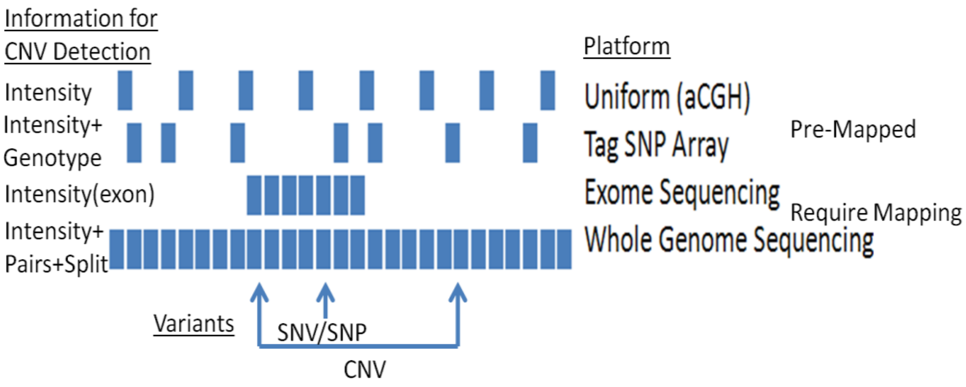
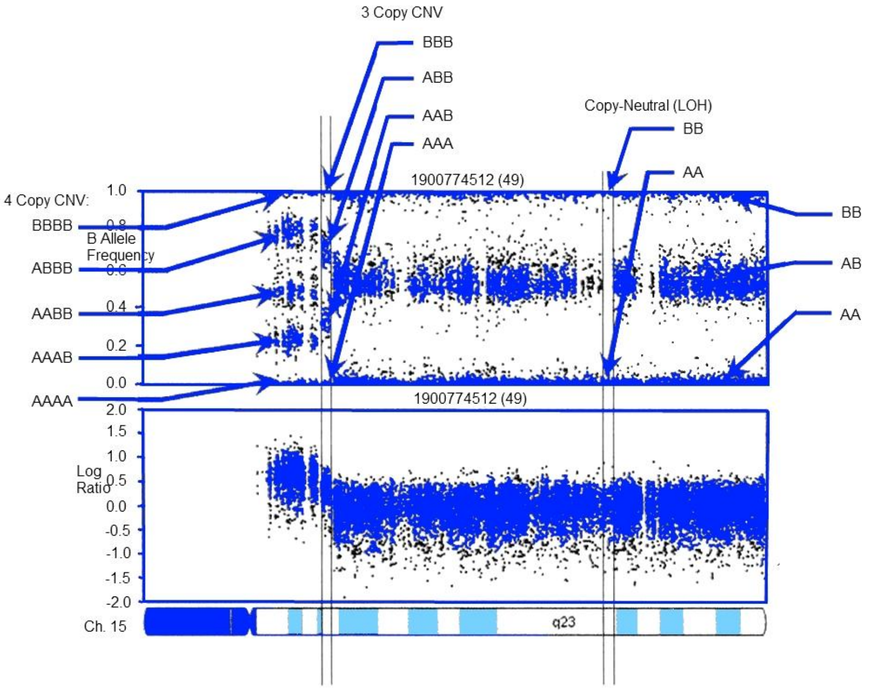


Figure 2, Figure 2. CNV detection in SNP-array data using PennCNV: Example Log R Ratio (LRR) and B Allele Freq (BAF) values for the chromosome 15 q-arm of an individual. Three normal chromosomal BAF genotype clusters (AA, AB, and BB genotypes) have LRR values around zero. The copy-neutral loss-of-heterozygosity (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[21]).


3 Copy CNV
BBB
ABB
AAB
AAA
Copy-Neutral (LOH)
BB
AA
1900774512 (49)
4 Copy CNV:
BBBB
B Allele
Frequency
ABBB
AABB
AAAB
AAAA
BB
AB
AA
1900774512 (49)
Log
Ratio
Ch. 15
q23