Published in final edited form as:
Genomics. 2009 January ; 93(1): 22–26. doi:10.1016/j.ygeno.2008.08.012.

GENETIC ASSOCIATION ANALYSIS OF COPY NUMBER VARIATION (CNVs) IN HUMAN DISEASE PATHOGENESIS

Iuliana Ionita-Laza[1], **Angela J. Rogers**[2], **Christoph Lange**[1,2], **Benjamin A. Raby**[2], and **Charles Lee**[3]

1 *Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA, 02115, USA*

2 *Channing Laboratory, the Division of Pulmonary and Critical Care Medicine, and the Center for Genomic Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*

3 *Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, 221 Longwood Avenue, Boston, MA 02115, USA*

Abstract

Structural genetic variation, including copy number variation (CNV), constitutes a substantial fraction of total genetic variability and the importance of structural genetic variants in modulating human disease is increasingly being recognized. Early successes in identifying disease-associated CNVs via a candidate gene approach mandate that future disease association studies need to include structural genetic variation. Such analyses should not rely on previously developed methodologies that were designed to evaluate single nucleotide polymorphisms (SNPs). Instead, development of novel technical, statistical, and epidemiologic methods will be necessary to optimally capture this newly-appreciated form of genetic variation in a meaningful manner.

Keywords

Copy number variation; CNV; structural genetic variation; disease association study; complex trait

"When all you have is a hammer, everything begins looking like nails."

Abraham Maslow (1908–1970)

Among the many important insights derived from completion of the Human Genome Project was the recognition of the abundance of single nucleotide polymorphisms (SNPs) as a major source of genetic variation, leading to speculation that the bulk of phenotypic variability in human populations is due to single base changes. As a result, intense efforts were made to develop high-throughput sequencing and SNP genotyping platforms, SNP databases, detailed linkage disequilibrium maps (through the International HapMap Project), and statistical methodologies for analyzing SNP genotype and haplotype data in mapping disease-susceptibility genes. Until recently, the overwhelming majority of gene-mapping studies have

Corresponding Author: Dr. Charles Lee, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, 221 Longwood Avenue, EBRC 404A, Boston, MA, USA 02115, Telephone: (617) 278-0031, Fax: (617) 264-6861, E-mail: clee@rics.bwh.harvard.edu.

focused exclusively on the role of SNPs in human diseases. Indeed, using population-based studies to identify genetic determinants of common disease, dozens of SNP-based susceptibility variants have been identified for human diseases as diverse as diabetes[1,2], macular degeneration[3,4], cancer[5,6], asthma[7] and Crohn disease [1,8]. However, studies over the past three years have resulted in increasing recognition of the critical role of structural genetic variation (most of which appear to be in the form of copy number variation) in modulating gene expression and disease phenotype. In fact, copy number variants (CNVs) are now known to be a prevalent form of common genetic variation and represent a substantial proportion of total genetic variability in human populations. Moreover, a few association studies have already demonstrated the importance of CNVs as disease-susceptibility variants, with specific CNVs found to confer differential risk to HIV infections [9], autoimmune disease [10–12], and asthma [13–16](Table 1). Recently, genome-wide surveys have demonstrated that rare CNVs altering genes in neurodevelopmental pathways are implicated in autism spectrum disorder [17] and schizophrenia [18]. It is therefore becoming increasingly clear that genetic studies of complex diseases must pay closer attention to the contribution of CNVs.

In contrast to the well-developed resources available for SNP association studies, we are still in the very early phases of incorporating structural genetic variation in genome-wide association studies. Nevertheless, we anticipate a burgeoning focus on structural genetic variation in human disease over the next few years, and foresee the development of many tools needed for such studies. In this commentary, we provide a very brief description of the presently known landscape of structural genetic variation, review recent successes in identifying CNVs associated with human diseases, and then address the current challenges of CNV association studies, including the limitations of current genotyping platforms and available statistical methods.

Presently known landscape of CNVs

Structural genetic variation refers to a class of genomic alterations of DNA that usually span more than 1000 bases (reviewed in Freeman *et al.* 2006 and Feuk *et al.* 2006)[19,20]. It includes quantitative (unbalanced) changes such as copy-number variants (CNVs), and less common balanced variations involving chromosomal inversions, insertions, and translocations. Here, we focus on CNVs, the most prevalent type of structural genetic variation.

Structural genetic variation has long been known to impact health, though until very recently this impact was thought to be limited to rare genomic disorders. A handful of Mendelian disorders, such as Williams-Beuren Syndrome (deletion at chromosome region 7q11.23) or Charcot-Marie Tooth neuropathy Type 1A (duplications of peripheral myelin protein-22 at chromosome region 17p11.2), are caused exclusively by recurrent DNA copy number changes at critical loci. However, with the realization of the existence of widespread common structural variation among otherwise healthy individuals [21–23], greater attention is now being focused on whether this type of genetic variation influences more common human diseases.

The current map of structural variation in the human genome is far from complete [24]. While several databases exist to catalog this newly-appreciated form of human genetic variation (notably the Database of Genomic Variants - http://projects.tcag.ca/variation/ and the Human Structural Variation Database - http://humanparalogy.gs.washington.edu/structuralvariation/), quality control is lacking, and studies have differed in technological approaches, precise boundary definition of CNVs, DNA quality, and even discrepancies in terminology [24]. Nevertheless, the latest compilation of data on structural genetic variation (The Database of Genomic Variants - November 29, 2007) from 46 different articles over the past three years indicate that as many as 4878 loci (comprised of 11,784 different CNV entries) have now been identified. We anticipate that our understanding of the location and extent of CNV in the human

genome will improve markedly in the next few years. Emerging technologies are more sensitive for detection of CNVs and provide more precise definition of boundaries (e.g. Perry et al. [25]). Undoubtedly, as a clearer map of human structural genetic variation emerges, we will begin to more comprehensively include this type of genetic variation in genome-wide association studies that attempt to elucidate the role of CNVs in human disease.

CNVs in health and disease

Several distinguishing features of CNVs support their role in disease pathogenesis. First, though less abundant than SNPs, it has been suggested that CNVs account for more nucleotide variation than do SNPs, on account of their sheer size [23] By spanning thousands of bases, CNVs often encompass (and can sometimes disrupt) functional DNA sequences. Second, there appears to be an enrichment of currently-known CNVs toward “environmental sensor” genes – i.e. genes that are not necessarily critical for early embryonic development, but rather help us to perceive and interact successfully with our ever-changing environment [22,23]. This includes enrichment for olfactory receptors, immune and inflammatory response genes, cell signaling and cell adhesion molecules, structural proteins, and ion channels. Third, like other forms of genetic variation, both purifying and adaptive natural selective pressures appear to have influenced the frequency distribution of selective CNVs, suggesting their functional significance [26–29]. Lastly, a recent comparison of the relative impact of SNPs and CNVs on gene expression noted that a substantial proportion (~18%) of gene expression variability was attributable to known CNVs greater than ~40 kb in size [30]. Notably, 53% of genes whose expression was influenced by CNVs had the corresponding CNV outside of the actual gene, suggesting that many CNVs could affect important regulatory sequences that are situated at a distance from the actual target gene.

Given these features, it is perhaps not surprising that early genetic association studies of known CNVs have quickly produced promising results. Presented in Table 1 are recent examples of copy-number-variable loci implicated in the pathogenesis of complex traits, where the association has been observed in at least two independent populations. These loci share several noteworthy features that may provide important insights into the role CNVs may play in complex diseases. First, the copy number frequencies for all five loci are high – greater than 10% in all cases - confirming that the allelic spectrum of disease-related CNVs is not restricted to rare variants. Second, with the exception of the CCL3L1 HIV/AIDS protective alleles, the genetic risk conferred by these variants is quite high (relative to SNPs, particularly in the context of polygenic, complex traits). Currently available data suggest that many CNVs confer greater disease risk than SNPs and in some cases these CNV-based disease susceptibility variants appear to increase risk by as much as 30%. Although we caution over interpretation of these early estimates (given that risk tends to be overestimated in initial studies due to the so-called “winner’s curse” [31,32]), these early returns do support an important role for CNVs in the genetic etiology of common diseases.

These early studies also suggest that copy-number variable loci may exhibit copy-number dependent genetic pleiotropy. We note that for two of the loci listed in Table 1, gains and losses are associated with distinct phenotypes (HIV and Rheumatoid Arthritis for CCL3L1; Crohn’s disease and Psoriasis for DEFB4) ([9,11,33,34]). These observations are reminiscent of neuropathies associated with copy number variation at the Peripheral Myelin Protein 22 (PMP22) locus, where PMP22 duplication confers Charcot-Marie-Tooth Disease (PubMed ID: 1677316) and PMP22 deletions cause hereditary neuropathy with liability to pressure palsies (also known as bulb diggers’ palsy) (PubMedID: 8422677). Whether this phenomenon will be observed at other copy-variable loci is unclear.

Another striking feature shared by the loci listed in Table 1 is that all are immune or inflammatory-related genes. Though certainly a function of the diseases studied in these surveys, this enrichment is consistent with the distribution of functional gene classes in CNV regions, where inflammatory and immune-related genes were among the most overrepresented [35].

It is important to recognize that all of these loci were identified using candidate gene approaches rather than hypothesis-free, genome-wide surveys. It is therefore unclear whether the above observations (relating to allele frequency distribution, effect sizes and functional class representation) will continue to hold as novel loci are identified through genome-wide association studies not predicated on prior biological knowledge. In one recent genome-wide CNV-association study on autism spectrum disorder (ASD) [17], 264 families (including 165 families with autistic children and 99 control families) were screened for *de novo* CNV mutations. In this study, the authors observed a disproportionate incidence of *de novo* mutations in families with ASD (12 deletions and 2 duplications among affected families compared to 2 gains among controls). This represented an approximate 3-fold increase in *de novo* mutation rate. All of the ASD-associated CNVs harbored at least one gene, several of which have been implicated in clinical contexts to overlap with autism. A similar study [18] showed the importance of rare CNVs at multiple sites in schizophrenia. In this study, the authors observed that novel (that is, not present in the Database of Genomic Variants) microdeletions and microduplications (> 100 kilobases) were present in 15% of schizophrenia cases, a frequency three times that in controls. Notably, mutations in cases disproportionately affected genes from signaling networks controlling neurodevelopment, including neuregulin and glutamate pathways.

The early successes described above suggest that there will be a sharp increase in the number of published CNV-association studies over the coming years. In anticipation of this, we stress that there still remain considerable technical, methodological, and analytical challenges related to CNV-based association studies that must be recognized and carefully addressed.

Technical Challenges in CNV studies

CNV-based association studies pose additional unique challenges, including choice of genotyping platform (for a recent review, please see Carter 2007 [36]) and DNA quality control. Three broad platform classes are currently available for genome-wide copy number surveys: (1) Large insert clone-based comparative genomic hybridization (CGH), where differentially-labeled test and reference samples compete for binding to DNA from large insert genomic clones - such as BACs (*e.g.*, Fiegler *et al.* 2006 [37]); (2) long, isothermic oligonucleotide-based CGH arrays [38] (where differentially-labeled test and reference samples compete for binding to 50 – 65mer oligonucleotides that are designed to have similar thermodynamic kinetics); and (3) SNP-based arrays [39] (one-sample arrays where intensity values derived from genotyping assays are used to infer copy number). Large insert clone-based CGH arrays have the highest signal-to-noise ratios, but are limited in their use to association studies due to their relatively low effective resolution (max. 15–35 kb). Conversely, SNP-based methods are optimized for high-throughput studies, and the high SNP density on most current arrays provide high resolution (~3 kb) in regions of the genome that are well represented. Furthermore, genome-wide SNP data are already being generated from genome-wide association studies (GWAS) across a variety of complex diseases, and thus it would be attractive to simply reanalyze these existing data for preliminary genome-wide CNV surveys. However, unfortunately, SNP-based arrays have the poorest signal-to-noise ratios (an order of magnitude worse than CGH), and it remains unclear whether the convenience of data availability will offset the high measurement error currently observed. Long oligonucleotide arrays offer both intermediate resolution and performance, and can be useful for high-resolution CNV detection,

validation and characterization. One adaptation of this platform, was used for the genome-wide CNV autism screen described by Sebat *et al* [17], and also for the schizophrenia study [18]. Newer platforms have now been developed that strategically include many probes (that aim to collectively detect CNVs, rather than SNPs) in a manner as to allow a single assay to effectively capture both SNP and CNV data (*e.g.*, the Affymetrix Human SNP 6.0 array and Illumina 1 million feature genotyping assay). Association studies using these platforms are ongoing and therefore preclude comment on their performance at this time.

All of the currently available platforms rely on efficient DNA hybridization, which can be significantly impaired by poor quality sample DNA. DNA sample quality is of particular concern in case-control studies where genomic DNA being used is collected from different sites and over long periods of time. As quantitative measures are sensitive to DNA degradation or contamination, spurious evidence of association may arise if sample ascertainment and DNA storage techniques differ between case and control samples. Similar bias can potentially be introduced in family-based studies if samples from probands and parents are collected under different conditions. Where possible, DNA samples from all sites should be handled uniformly, and an assessment of the DNA quality (including OD measurements), estimates of sample purity, DNA degradation and DNA concentration should be performed prior to genotyping.

Statistical challenges in analysis of CNV associations with human disease

Genetic epidemiology of CNVs is still in its infancy - so too are the statistical methods for the analysis of CNV association with disease. As discussed above, genotyping platforms vary in their signal-to-noise ratio and also in their ability to define precise CNV genotypes (i.e. discrete copy-number). Current CNV typing technologies produce quantitative measures meant to reflect total DNA amount in a given sample. In contrast to SNP genotyping, oftentimes the distribution of these measurements is continuous, making it difficult to accurately estimate DNA copy number. Currently only a small fraction of the known CNVs are genotypable.

Two main statistical methodologies can be employed for CNV analyses, and they differ in their need for precise CNV quantification. The first involves a two-step procedure, based on first inferring the underlying genotype and then performing a regular test of association [40,41]. Because it depends on “genotypable” markers, this class of statistical methodologies is currently only applicable to a limited number of CNVs. Also, it is not immediately clear how the uncertainty of CNV genotype calling should be incorporated in the analysis; when raw measurements show a continuous distribution, forcibly classifying them into discrete copy number classes (*e.g.* gain, no change, or loss) may result in the loss of substantial information and statistical power relative to the raw measurements [42] (see also Figure 1). The second methodology bypasses the genotype calling step and instead, directly analyzes the intensity measurements which are thought to reflect the true underlying copy number. This strategy has been advocated in several recent papers [30,42].

For case-control designs, classical statistical methods (e.g. parametric (t-tests), non-parametric (Mann-Whitney U test) etc.) can be employed. Also, some of the methods that are already available for SNP genotype data can potentially be adapted to the analysis of signal intensity data (e.g. population stratification methods such as Eigenstrat, [43]). For family-based study designs, a method extending family-based tests (FBATs) [44,45] to deal with copy number variation has recently been proposed [46]. Like the FBAT method for SNP data, CNV-FBAT is based on the covariance between offspring trait and offspring CNV data (adjusted for the parental CNV data), where the CNV data here are represented by the normalized signal intensity measurements. The robustness properties of the genotype FBAT-approach are maintained and previously developed FBAT extensions (including FBATs for time-to-onset, multivariate FBATs, and FBAT-testing strategies) can be directly transferred to the analysis

of CNVs [44]. This methodology has the advantage of using as much data as possible, without affecting power in a significant way. However, while signal intensity-based CNV data provide valuable information and can be used directly to perform association tests, knowing the true underlying genotype is important for validating the association signal and providing further insight into the biological mechanisms by which a CNV influences the disease.

Failure to correct for biases in data collection or processing prior to performing the association test may result in spurious associations with either of the statistical approaches. Therefore, it is mandatory that disease associations be verified using alternate technologies and/or independent datasets. This potential problem may be accentuated with the case-control design, which depends on selection of a suitable control sample. As discussed above, the control sample needs to come from a comparable population, with nearly identical DNA quality, preparation, and handling among samples in the case and control populations. Family-based designs can alleviate some of these problems, as the parents and other family members of the proband act as well-matched controls. However, care needs to be taken that samples from family members are collected and evaluated under similar conditions as the proband. A further advantage of family-based study designs is their potential for discriminating between *de novo* and inherited CNVs, once an association between a CNV and a phenotype has been established.

Epidemiologic challenges in the design of CNV-based association studies

The novelty of studying CNV should not obscure the need for meticulous attention to epidemiologic study design. CNV association studies should employ the same rigorous standards that have been used in conventional SNP-based genetic epidemiology studies. These include assurance of adequate sample size to detect modest genetic effects, appropriate adjustment of significance thresholds to adjust for multiple comparison testing, provision of evidence for reproducible association in independent populations, and rigorous assessment and adjustment for population stratification. This latter point warrants elaboration. Like SNPs, most common CNVs are shared between populations and follow frequency distributions reminiscent of other forms of genetic variation [28]. However, like SNPs, some CNVs demonstrate notable between-population differences in allele frequency distribution (representing ~11% of the variance in between population differences) [28]. As a consequence, CNV-association studies, like SNP-association studies, are susceptible to bias from population stratification, whereby spurious genetic association could be observed between marker and trait simply due to differing ancestral composition (and hence genotype frequency distribution) between cases and controls [28,47]. There are several methods for addressing population stratification, including stringent sampling of cases and controls from homogenous ancestral groups, use of family-based designs, and via analytic methods (including sequential screening for and quantifying population stratification using random sets of markers with subsequent adjustment of association test statistics by the degree of observed stratification [48]).

Conclusions

The study of structural genetic variation in human diseases is a new and rapidly evolving field. The main limitations of this field of study relate to the lack of technological and statistical tools dedicated to these efforts, and reliance on the “hammers” developed for the study of SNPs. Over the next few years, a much more thorough understanding of the extent and precise location of copy number variation will likely be available as new platforms become available to accurately capture CNV information. This will allow us to more comprehensively understand the role of these genetic variants in the pathogenesis of human diseases. However, in addition to the development of better CNV genotyping platforms, we stress that rigorous attention to study design and statistical analysis is critical so as not to relive past experiences of early disease association studies that yielded “unreplicable” and all too often false-positive results.

Whenever possible, initial reports of CNV-disease association should include independent evidence of replication in other studies and populations. Generating such data will frequently require collaboration between research groups, and may include sharing of DNA samples given the current technical challenges of CNV genotyping and operator-dependence of quantitative genotyping assays like qPCR. Despite the many obstacles yet to be overcome, we foresee CNVs quickly taking their place alongside SNPs in genetic epidemiology studies. Once identified, these loci can then be evaluated experimentally using animal models that recapitulate the disease-associated molecular defect (*e.g.*, knock-out mice for CNV losses and over-expressing transgenic models for CNV gains) and the development of specific molecular therapeutics, ultimately leading to novel therapies for our most common diseases.

References

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78. [PubMed: 17554300]
2. Saxena R, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331–6. [PubMed: 17463246]
3. Edwards AO, et al. Complement factor H polymorphism and age-related macular degeneration. Science 2005;308:421–4. [PubMed: 15761121]
4. Klein RJ, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–9. [PubMed: 15761122]
5. Easton DF, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–93. [PubMed: 17529967]
6. Hunter DJ, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–4. [PubMed: 17529973]
7. Moffatt MF, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448:470–3. [PubMed: 17611496]
8. Duerr RH, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 2006;314:1461–3. [PubMed: 17068223]
9. Gonzalez E, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434–40. [PubMed: 15637236]
10. Fanciulli M, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 2007;39:721–3. [PubMed: 17529978]
11. McKinney C, et al. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis. 2007
12. Yang Y, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 2007;80:1037–54. [PubMed: 17503323]
13. Brasch-Andersen C, et al. Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. Hum Mutat 2004;24:208–14. [PubMed: 15300848]
14. Ivaschenko TE, Sideleva OG, Baranov VS. Glutathione- S-transferase micro and theta gene polymorphisms as new risk factors of atopic bronchial asthma. J Mol Med 2002;80:39–43. [PubMed: 11862323]
15. Palmer CN, et al. Glutathione S-transferase M1 and P1 genotype, passive smoking, and peak expiratory flow in asthma. Pediatrics 2006;118:710–6. [PubMed: 16882827]
16. Piirila P, et al. Glutathione S-transferase genotypes and allergic responses to diisocyanate exposure. Pharmacogenetics 2001;11:437–45. [PubMed: 11470996]
17. Sebat J, et al. Strong association of de novo copy number mutations with autism. Science 2007;316:445–9. [PubMed: 17363630]
18. Walsh T, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 2008;320:539–543. [PubMed: 18369103]

19. Freeman JL, et al. Copy number variation: new insights in genome diversity. Genome Res 2006;16:949–61. [PubMed: 16809666]
20. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet 2006;7:85–97. [PubMed: 16418744]
21. Iafrate AJ, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–51. [PubMed: 15286789]
22. Sebat J, et al. Large-scale copy number polymorphism in the human genome. Science 2004;305:525–8. [PubMed: 15273396]
23. Tuzun E, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727–32. [PubMed: 15895083]
24. Scherer SW, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet 2007;39:S7–15. [PubMed: 17597783]
25. Perry GH, et al. The fine-scale and complex architecture of human copy number variation. Am J Hum Genet. (In Press)
26. Nguyen DQ, Webber C, Ponting CP. Bias of selection on human copy-number variants. PLoS Genet 2006;2:e20. [PubMed: 16482228]
27. Perry GH, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007
28. Redon R, et al. Global variation in copy number in the human genome. Nature 2006;444:444–54. [PubMed: 17122850]
29. Stefansson H, et al. A common inversion under selection in Europeans. Nat Genet 2005;37:129–37. [PubMed: 15654335]
30. Stranger BE, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007;315:848–53. [PubMed: 17289997]
31. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306–9. [PubMed: 11600885]
32. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177–82. [PubMed: 12524541]
33. Fellermann K, et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 2006;79:439–48. [PubMed: 16909382]
34. Hollox EJ, et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 2008;40:23–25. [PubMed: 18059266]
35. Cooper GM, Nickerson DA, Eichler EE. Mutational and selective effects on copy-number variants in the human genome. Nat Genet 2007;39:S22–9. [PubMed: 17597777]
36. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 2007;39:S16–21. [PubMed: 17597776]
37. Fiegler H, et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res 2006;16:1566–74. [PubMed: 17122085]
38. Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. J Clin Pathol 2004;57:644–6. [PubMed: 15166273]
39. Komura D, et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res 2006;16:1575–84. [PubMed: 17122084]
40. Kohler JR, Cutler DJ. Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. Am J Hum Genet 2007;81:684–99. [PubMed: 17846995]
41. Kosta K, et al. A Bayesian approach to copy-number-polymorphism analysis in nuclear pedigrees. Am J Hum Genet 2007;81:808–12. [PubMed: 17847005]
42. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet 2007;39:S37–42. [PubMed: 17597780]
43. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9. [PubMed: 16862161]

44. Lange C, Laird NM. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. Genet Epidemiol 2002;23:165–80. [PubMed: 12214309]
45. Lange C, et al. A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. Am J Hum Genet 2003;73:613.(abstract)
46. Ionita-Laza I, et al. On the analysis of copy-number variations in genome-wide association studies: A translation of the family-based association test. Genetic Epidemiology. 2008in press
47. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet 2007;39:S30–6. [PubMed: 17597779]
48. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet 2004;36:512–7. [PubMed: 15052271]
49. Townson JR, Barcellos LF, Nibbs RJ. Gene copy number regulates the production of the human chemokine CCL3-L1. Eur J Immunol 2002;32:3016–26. [PubMed: 12355456]
50. Nibbs RJ, Yang J, Landau NR, Mao JH, Graham GJ. LD78beta, a non-allelic variant of human MIP-1alpha (LD78alpha), has enhanced receptor interactions and potent HIV suppressive activity. J Biol Chem 1999;274:17478–83. [PubMed: 10364178]
51. Aitman TJ, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 2006;439:851–5. [PubMed: 16482158]
52. Olsen ML, et al. C4A gene deletion and HLA associations in black Americans with systemic lupus erythematosus. Immunogenetics 1989;30:27–33. [PubMed: 2568334]
53. Childhood Asthma Management Program Research Group. The Childhood Asthma Management Program (CAMP): design, rationale, and methods. Control Clin Trials 1999;20:91–120. [PubMed: 10027502]
54. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 2004;99:96–104.

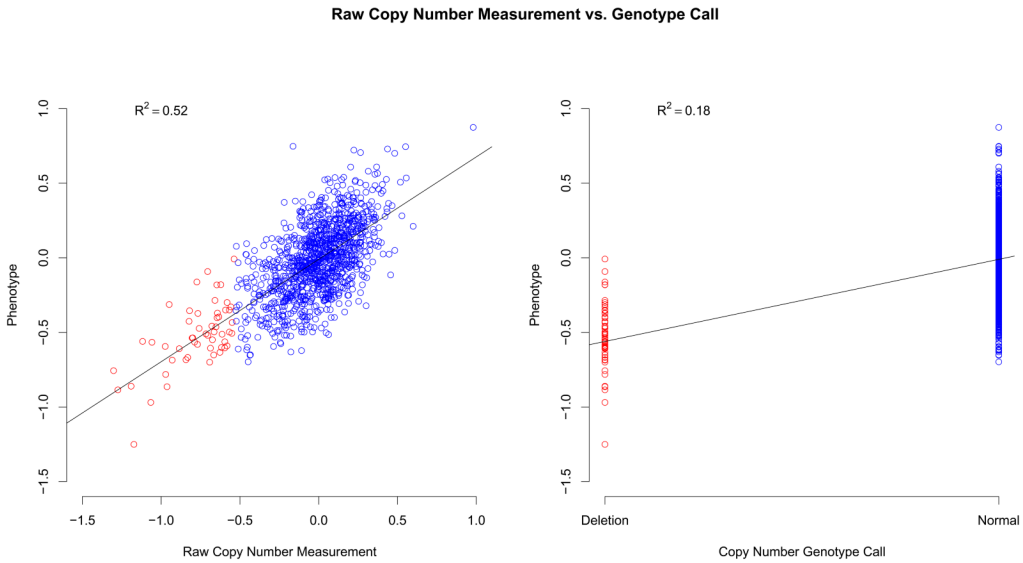

Figure 1. Raw copy number measurements vs. CNV calls for a CNV showing a continuous distribution

Shown in this figure, on the left side, is the association between a normally distributed, simulated phenotype and the intensity measurements at a single SNP position within a known copy-number variable region on chromosome 21 using a dataset of approximately 1200 individuals (CAMP study [53]). On the right side, the association between the same phenotype and the CNV calls (loss/no change) is shown. Losses were detected using a local false discovery rate (locFDR) approach [54], applied to the intensity measurements. As can be observed, for CNVs showing a continuous intensity distribution, forcibly classifying the raw measurements into discrete calls may result in loss of power compared to the original measurements, as illustrated by the drop in the R^2 value (the square of the correlation coefficient).

Table 1

Replicated associations of DNA copy number variants with common complex disease

Locus	CNV Frequency	Clinical phenotype	CNV type	Risk estimate (Odds Ratio)	Comments
CCL3L1 [9,11]	10–20%	HIV/AIDS susceptibility [9]	Deletion	0.67–0.90	CCL3L1 inhibits HIV cellular entry [49] Higher CCL3L1 number increases CCL3L1 expression [49]
		Rheumatoid Arthritis [11]	Gain: >2	1.34	
FCGR3B [10]	Deletion: ~25% Gain: ~15%	Systemic Autoimmune disease	Deletion	1.58–2.56*	CNV associated with glomerulonephritis in rats and humans [51]
C4[12]	~40%	Systemic Lupus Erythematosis	Deletion	Absence: 5.27 Carrier: 1.61 Gains: 0.57	> 75% of C4 or C1 deletion carriers have SLE-like disease [12] Strongest SLE genetic risk factor thus far in blacks [52]
DEFB4 [33,34]	2–12 copies (median 4)	Colonic Crohn Disease [33]	Loss: <4 copies	3.06	↓ number associated with ↓ mucosal gene expression. [33]
		Psoriasis [34]	Gain: >5 copies	1.69	
GSTM1 [13–16]	Up to 50%	Asthma, lung function, allergic response	Deletion	1.59–1.89	Potent antioxidant. Deletion related to many adverse asthma-related outcomes (see text).

* FCGR3B demonstrates phenotypic pleotropy: OR for lupus 2.21; for Wegener's Granulomatosis 1.58–2.46; for microscopic polyangiitis 2.56.