PERSPECTIVE

nature genetics

Copy-number variation and association studies of human disease

Steven A McCarroll & David M Altshuler

The central goal of human genetics is to understand the inherited basis of human variation in phenotypes, elucidating human physiology, evolution and disease. Rare mutations have been found underlying two thousand mendelian diseases; more recently, it has become possible to assess systematically the contribution of common SNPs to complex disease. The known role of copy-number alterations in sporadic genomic disorders, combined with emerging information about inherited copy-number variation, indicate the importance of systematically assessing copy-number variants (CNVs), including common copy-number polymorphisms (CNPs), in disease. Here we discuss evidence that CNVs affect phenotypes, directions for basic knowledge to support clinical study of CNVs, the challenge of genotyping CNPs in clinical cohorts, the use of SNPs as markers for CNPs and statistical challenges in testing CNVs for association with disease. Critical needs are high-resolution maps of common CNPs and techniques that accurately determine the allelic state of affected individuals.

Empirical evidence that CNVs are associated with phenotypes

The first evidence that copy-number alterations can influence human phenotypes came from sporadic diseases, termed 'genomic disorders', caused by *de novo* structural alterations[1]. The number of genomic disorders has grown, with several dozen reported to date[2]. In addition to such sporadic diseases, inherited CNVs have been found to underlie mendelian diseases in several families[3–5]. Nonetheless, CNVs have been implicated in only a few percent of the 2,000 or more mendelian diseases so far explained at a molecular level.

Little is known about the genetic basis of common, complex phenotypes, and thus it would be premature to predict the relative proportion of complex disease explained by SNPs and CNVs. In principle, complex disease might be more susceptible to 'soft' forms of variation — such as variation in noncoding sequences and copy number — which alter gene dose without abolishing gene function. Common CNPs have been reported to be associated with several complex disease phenotypes, including HIV acquisition and progression[6], lupus glomerulonephritis[7] and three systemic autoimmune diseases: systemic lupus erythematosus, microscopic polyangiitis and Wegener's granulomatosis[8,9]. A recent study of gene expression variation as a model complex phenotype measured the fraction of gene expression 'traits' that could be associated with either SNPs or CNVs; in this study, SNP genotypes and CNV measurements were associated with 83% and 18% of those gene expression traits for which statistically significant associations were found[10]. This may still underestimate the role of CNVs, given the greater completeness and accuracy with which SNPs can be queried at present.

Technical issues in assessing CNVs for a role in disease

The power to discover a relationship between DNA variation and phenotype is limited by the sensitivity and accuracy with which that DNA variation is measured in each individual. (Accuracy in measuring phenotype, environment and behavior are equally or more important; these challenges, which are not specific to CNV studies, are beyond the scope of this review.) To the extent that the precise allelic state of any DNA variant is not well measured, power declines. But although this issue has been the subject of extensive discussion in the literature on SNP association studies[11], little has been written about the extent to which current attempts to measure copy number provide precise and accurate measures of the underlying DNA variation in each individual.

Insufficient data have been collected at a sequence level to estimate the correlation between quantitative measures of CNVs by existing techniques and the true allelic state of each CNV in each individual. However, there are reasons to believe that the correlation is low — much lower, for example, than that offered by current SNP genotyping products for the underlying SNP variation in the genome[12]. This difference is due fundamentally to the greater challenge of measuring multibase, often multiallelic variants compared with single-base, diallelic SNPs. The specific challenge of genotyping CNVs is discussed in a later section.

Enabling core knowledge about CNVs for association studies

An indispensable starting point for association studies is basic knowledge about the genetic variations that are present in the human population — their alleles, their frequencies, their precise locations. SNP databases developed through a series of phases: first, rapid growth in methods to detect the locations of putative SNPs; second, calibration and standardization of discovery methods to maximize sensitivity and

The authors are in the Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA; and Department of Molecular Biology and the Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. David Altshuler is also in the Department of Medicine at Massachusetts General Hospital and the Departments of Genetics and Medicine, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: smccarro@broad.mit.edu

Published online 27 June 2007; doi:10.1038/ng2080

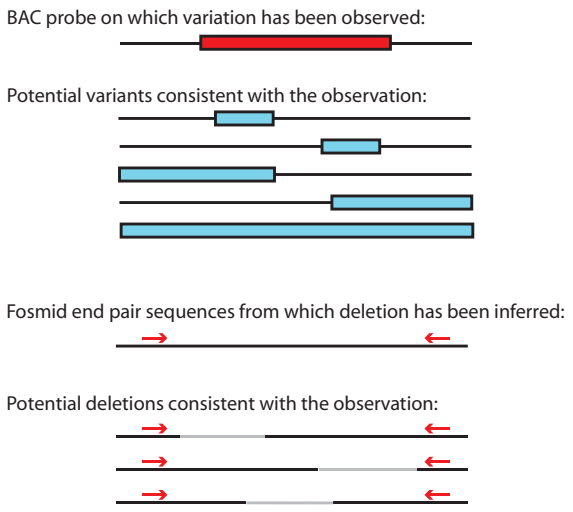

Figure 1 Many potential CNV locations are consistent with the coordinates of a reported CNV-containing region (CNVR). An investigator will typically have to perform extensive experiments to find the CNV within a CNVR. An ongoing project to map structural variants at the sequence level[28] will provide important enabling knowledge for clinical genetic studies.

minimize false positives[13]; third, accurate genotyping of large numbers of SNPs to validate (or invalidate) SNPs and characterize their properties[14–16]; and fourth, assessment of the sensitivity of the resulting map in comparison to systematic resequencing data[16]. The resulting resource enables researchers and technology companies to design assays for any given SNP of interest (or a genome-wide collection) and to assess the relationship between any given SNP and others that may or may not have been typed.

Pioneering genome-wide surveys of CNVs[17–26] and databases holding the results of those studies[27] are an important initial step toward a comprehensive database of CNVs, but much work lies ahead. Most reported CNV locations actually correspond to the locations of CNV-containing regions (CNVRs), generally the genomic coordinates of a BAC probe, set of oligonucleotide probes, or fosmid from which a variant was discovered. A reported CNVR is consistent with a large number of potential variants (**Fig. 1**); of importance for the design of assays, seldom is it known which precise locus or gene within the CNVR is actually affected. An important step toward enabling knowledge of CNP locations is an ongoing effort to sequence fosmid clones containing structurally variant haplotypes[28]. Until the locations of CNVs within reported CNVRs are known, researchers interested in studying a reported CNV in clinical samples must first perform experiments to find the CNV(s) within the reported CNVR. Ultimately, the utility of CNV databases will be enhanced by data on the validation state of each putative CNV (to avoid wasting resources on false positive CNVs) and on the frequency of each allele in different populations (to estimate statistical power when designing association studies), and by a map of the linkage disequilibrium (or lack thereof) with nearby SNPs that may be easier to genotype or may already have been assessed.

Genotyping CNVs in disease association studies

Disease association studies rely on accurate genotypes. Most CNV studies to date have been discovery studies (generating lists of regions that contain CNVs) rather than association studies (assessing the correlation between phenotype and genotype). The underlying problems in CNV discovery and CNV genotyping are different. A discovery study begins with a null hypothesis that no variation exists at a locus and assesses whether the evidence for variation exceeds a genome-wide significance threshold; a high false-negative rate (failure to discover variants) is tolerated in order to preserve a low false-positive rate[29]. An association study tests a null hypothesis that variation is not associated with phenotype; all forms of misclassification (both over- and underascertainment of altered copy-number levels) are problematic, and all levels of copy number must be distinguished. This is a much more exacting requirement: for example, of some 1,500 CNVs that were identified in one recent genome-wide survey, only 70 common, diallelic CNVs yielded genotypes of the quality that could be used for linkage disequilibrium analysis[26]. Indeed, of the thousands of CNV-containing regions that have been identified in the literature to date, only a few percent have been genotyped in available reference samples (**Fig. 2**).

The development of assays for accurately typing CNVs in clinical cohorts has required enormous effort in CNV-disease association studies to date[6–9] and is one of the most pressing needs in CNV research. Although it is not yet clear what technology will be used, it is critical that any assay be reproducible in other labs. Standards for publishing a CNV-disease association should include genotypes for a publicly available set of reference samples, which can be used by other labs to develop additional assays and to assess the original assay.

Copy-number measurements versus copy-number genotypes

At any given nucleotide, biological copy numbers are integers. The more precise and localized a measurement of copy number, the more its distribution in a population shows a discrete distribution reflecting the underlying integer distribution of copy numbers (for example, 0, 1, 2; or 2, 3, 4; or even 2, 3, 4, 5, 6) (**Fig. 3**). Frequently, though, copy-number measurements seem to be continuously distributed across a population (**Fig. 3c**). The factors which cause imprecision in copy-number measurement can be divided into two categories. Measurement imprecision refers to the noise inherent in making any measurement. Spatial imprecision occurs when an assay aggregates information across a large region into a single measurement.

Surveys of copy-number variation have further summarized copy-number measurements into discrete values of 'gain' or 'loss' in each sample; although these assessments are sometimes referred to as 'genotypes', inspection of the underlying data often shows that these discrete distinctions are not reflected in the underlying distribution of measurements (**Fig. 3c**,**e**). Summarizing raw copy-number measurements into such 'calls' may lose information present in the original measurements, and is of uncertain relationship to the true genotype (**Fig. 3e**,**g**).

Until approaches for genome-wide CNP genotyping mature, a placeholder strategy may be to rely on raw hybridization measurements as an approximation to an unknown, underlying genotype. This approach was used in a recent study of CNPs and gene expression[10] that used copy number data from an immediately preceding study[26]; the analysis dispensed with the CNV 'calls' from the previous study, instead using the raw hybridization measurements for association analysis[10,26] (**Fig. 3**). The paucity of effective CNP genotypes means that techniques and algorithms for making genotype calls are a critical need in CNP disease research; until such approaches mature, raw measurements may be the preferred basis for a preliminary association analysis.

Using SNPs as markers for CNPs

Given the technical challenges in finding and typing CNPs, and the early stage of basic knowledge about their locations and molecular structures, an appealing strategy might be to rely on more-easily-typed SNPs to serve as markers by linkage disequilibrium for common variants

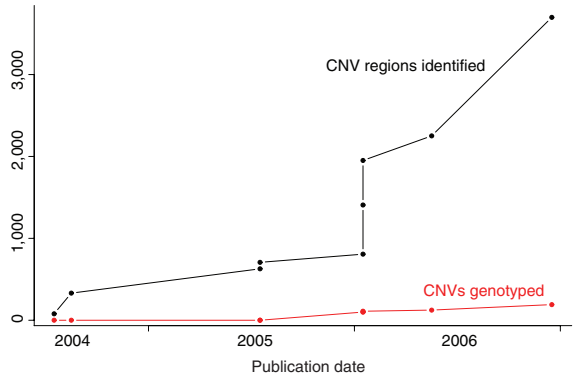

Figure 2 Although the number of reported copy-number-variable regions has increased dramatically, only a few percent of CNVs have been successfully genotyped. An important step in association study of any CNV will be the development of a genotyping assay that accurately determines the allelic state of every individual in a cohort.

throughout the genome. Linkage disequilibrium–based approaches utilize the observation that the human recombination rate is (i) low relative to the typical age of alleles in the human population and (ii) clustered into hotspots across the genome[30]. These features mean that ancestral variants (whatever their molecular nature) segregate in the population on haplotypes, are correlated with one another and thus can be 'tagged' by a reduced set of SNPs[31]. Because such linkage disequilibrium–based approaches require neither a priori identification of all variants nor technology for typing every variant individually, they might address the limitations of current knowledge and genotyping technology in the CNP field.

To assess a specific CNP through linkage disequilibrium, one would genotype the CNP in the HapMap (or other reference) samples and assess whether nearby SNPs were able to capture the CNP through linkage disequilibrium; if so, one would then type those SNPs in affected cohorts as a proxy for the CNP. To analyze a genomic region, one would select a dense set of SNPs sufficient to capture almost all common, ancestral polymorphisms through linkage disequilibrium[11] and test them for association with disease. On a genome-wide scale, one would presumably use commercial whole-genome SNP genotyping products. In all cases, positive association (if found) could be due to a CNV or to anything else in linkage disequilibrium with the associated SNP — possibilities that would be assessed by directed resequencing, copy-number analysis and additional genotyping in following up any initial association.

The performance of linkage disequilibrium–based approaches will depend on the strength and generality of linkage disequilibrium between CNPs and SNPs. Using available SNP data and PCR-based genotyping of deletion polymorphisms, initial studies found that deletion polymorphisms are generally ancestral and are tagged by SNPs[22,23]. A subsequent study of the linkage-disequilibrium properties of CNPs in the genome's segmental-duplication-rich regions found that copy-number measurements from such CNPs were less well captured by HapMap SNPs[24]; a more recent study of 70 genotyped CNPs found that the CNPs showed appreciable linkage disequilibrium with SNPs, but were less well tagged than frequency-matched SNPs were[26]. The extent of linkage disequilibrium between SNPs and CNPs remains unclear, for two reasons. First, assessing linkage disequilibrium around CNPs requires accurate genotyping of a large and representative collection of CNPs in samples with dense SNP genotypes — and yet accurate genotypes exist for only a small and nonrandom collection of CNPs (**Fig. 2**). Second, regions rich in segmental duplications contain almost half of all reported CNPs[19,24,26], but contain a density of validated SNPs (that could serve as potential tags) much lower than that of the rest of the genome[24].

Integrated association studies for SNPs and CNPs

Many genome-wide SNP association studies, each involving hundreds to thousands of affected individuals, are underway. The raw intensity data generated during SNP genotyping can be mined for copy-number information[32–35], making such studies a potential source of data for CNP-disease association studies. However, several factors limit the utility of previous generations of SNP arrays for this purpose. Most important is coverage: because common CNPs cause SNP genotyping assays to fail Hardy-Weinberg and mendelian inheritance checks, genomic regions harboring common CNPs had been filtered out (partially or completely) of commercial whole-genome SNP array platforms during the selection of high-performance SNP assays. Another limitation is technical: because SNP assays are optimized for allelic discrimination rather than copy-number measurement, the copy-number measurements they provide are noisy, with the result that only large variants are typically detected. Commercial SNP arrays are used to find the large copy-number alterations typical of cancer[32–35], but have not to date been used to perform association studies for germline CNPs, and seem to detect many more rare CNVs than common CNPs[26].

Ideally, every DNA sample would be simultaneously queried for SNPs and CNVs in a single, integrated analysis. We have been working with collaborators to develop hybrid oligonucleotide arrays that contain both SNP allele-discrimination probes and dedicated 'copy-number probes' — probes whose sequences have been optimized for copy-number quantification by (i) designing them to nonpolymorphic sequences, (ii) selecting sequence features predictive of technical efficacy and (iii) empirically assessing responsiveness in screening experiments (**Fig. 3d**,**f**,**h**). Such hybrid arrays (or some other technological solution) offer the potential for integrated association studies in which SNP and copy-number variation are considered together. Moreover, as databases of CNPs and SNPs become ever more complete, the content of such arrays should similarly approach completeness.

Testing the disease association of common CNPs

Once an accurate and complete set of CNV measurements is obtained in a sample, there are few unprecedented statistical challenges to the assessment of association with disease. As with SNPs, a key dividing line is whether the statistical test involves common variants or a collection of individually rare events.

For common CNPs, statistical tests will involve a straightforward comparison of allele frequencies (or of diploid genotype frequencies): between affected individuals and controls in a population cohort; between transmitted and untransmitted chromosomes in families with affected offspring; or between affected and unaffected siblings. Most successfully genotyped CNPs seem to be diallelic, showing 2 or 3 diploid copy-number classes and therefore most likely representing two underlying alleles[23,26]. Such variants are readily incorporated into current frameworks for SNP association testing; in fact, the copy-number classes could be subjected to the same quality-control tests (mendelian inheritance, Hardy-Weinberg equilibrium) used to ensure the quality of SNP genotypes. Such CNPs could for practical reasons be recoded as SNP genotypes (for example, 'AA' for zero copies, 'AC' for one copy, 'CC' for two copies) and thereby benefit from the software and analytical approaches already available for SNP-based analyses, including correcting for population stratification (discussed below) and scrutinizing a

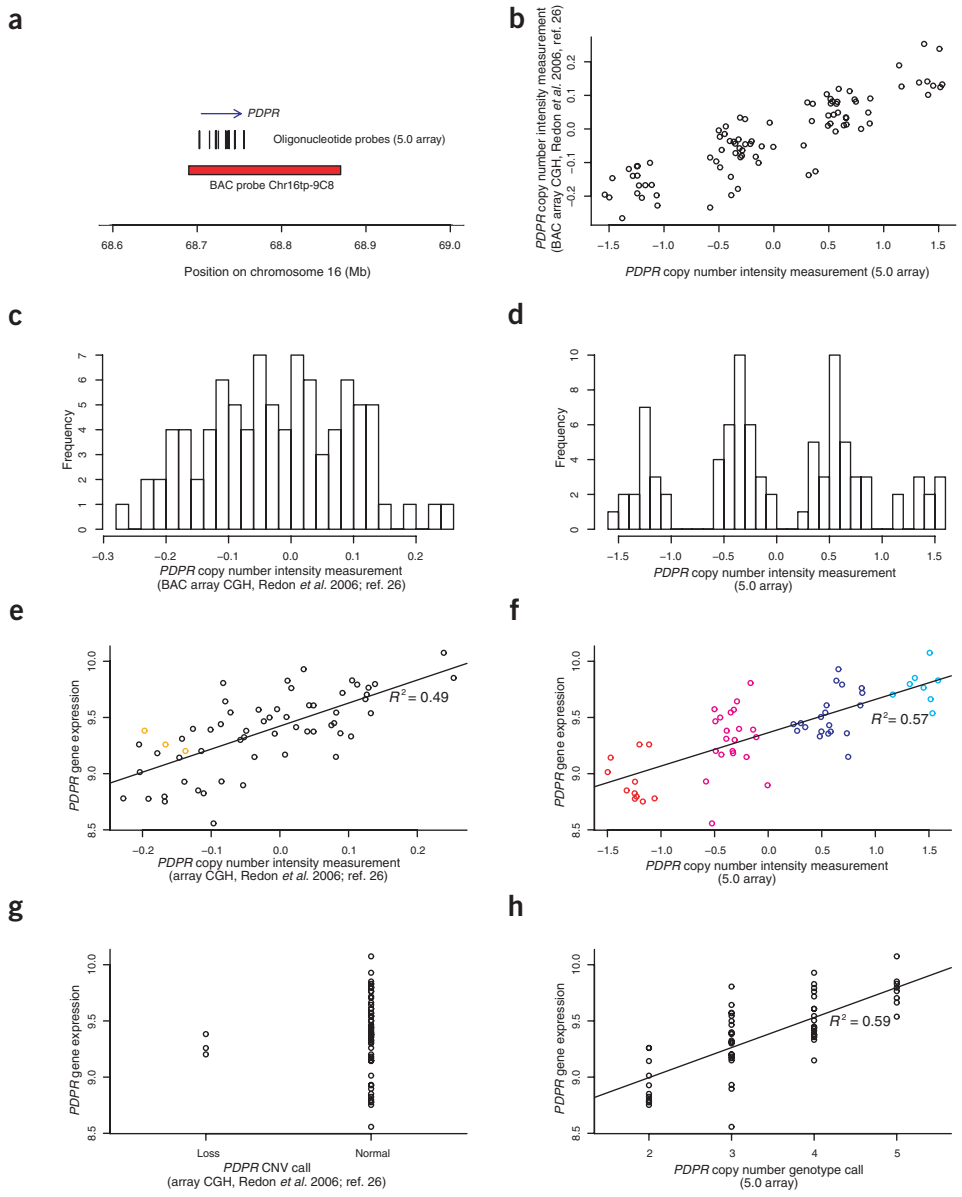

Figure 3 Using copy-number measurements and copy-number genotypes in association studies. (**a**) A common CNP containing a gene encoding pyruvate dehydrogenase phosphatase regulatory protein (*PDPR*; blue arrow) interrogated by a BAC probe (Chr16tp-9C8, red rectangle) on a BAC array-CGH platform[26], and by a series of oligonucleotide probes (vertical line segments) on an oligonucleotide platform (Affymetrix GenomeWide 5.0). (**b**–**d**) Copy-number measurements for the two platforms across the same set of samples (HapMap CEU sample of individuals with European ancestry) are correlated (**b**), confirming that they interrogate the same CNP. Measurements on the BAC array-CGH platform[26] show a continuous distribution (**c**), whereas measurements on the oligonucleotide platform show a discrete distribution (**d**). (**e**,**f**) Association analysis using raw intensity measurements. *PDPR* gene expression is found to be associated with *PDPR* copy number using raw measurements of copy number on both platforms. Colors indicate the discrete genotype 'calls' on each platform (not used in this analysis, but used in the analysis in panels **g**,**h**). The association in **e** was discovered by Stranger *et al.* (2007)[10]. (**g**,**h**) Association analysis using discrete genotype 'calls'. Where raw measurements show a continuous distribution (**c**,**e**,**g**), hardening of raw measurements into discrete 'call' loses information that was present in the original measurements, with the result that association with phenotype is no longer detected. Where raw measurements show a discrete distribution (**d**,**f**,**h**), conversion of raw measurements into genotypes can increase the correlation with phenotype, though the primary benefit may simply be greater clarity about the distribution of genetic variation and its relationship to phenotype.

genome-wide study for *P*-value inflation.

Some CNPs seem to involve more than three copy-number classes, and therefore more than two copy-number alleles (**Fig. 3**). Nineteen such loci were identified in a recent genome-wide CNV survey[26]. A related class of CNPs appears to harbor both deletion and duplication alleles[24,26]. Notably, the common CNPs reported to be associated with HIV progression and autoimmune phenotypes are multiallelic[6–9]. For the population-based analyses in those studies, researchers used a variety of techniques to test for disease association, including (i) reducing the copy-number genotype to a binary class (for example, >4 versus <4 copies), then performing a chi-squared analysis on the distribution of disease status between these two groups[6]; (ii) a logistic regression analysis, with copy number as an explanatory variable and age and gender as covariates[7]; and (iii) nonparametric tests of the null hypothesis that affected individuals and controls were drawn from the same distribution of copy numbers[7]. Family-based analyses — which are favored by many researchers because they are more robust to population stratification (discussed below) — will also need to be generalized to address multiallelic CNPs and continuously distributed copy-number measurements.

Testing the disease association of rare CNVs

For rare variants, association analysis is more challenging, as it is less constrained: there are many potential ways to group a collection of unique events, and thus more degrees of freedom. When copy-number ascertainment was limited to large, microscopically visible (and therefore usually functional) variants, such variants were generally assumed to be causative (although the specific gene involved is, conversely, very imprecisely localized). The new ability to detect smaller, submicroscopic CNVs — hundreds of which may be present in any one individual, and the vast majority of which are benign — requires statistically well

founded assessment of their association with disease.

As submicroscopic CNVs cannot be assumed to have functional consequence, it is critical to search for them in affected individuals and controls with equal rigor, and to use a statistical framework to determine whether rearrangements are truly more common in the affected. It is critical that CNVs not be discovered in a set of cases and then the specific variants that were found queried in controls; such an approach is subject to 'ascertainment bias' and is statistically unsound. Given the existence of hundreds of rare CNVs with apparent frequencies of less than one percent, even in a well designed study it will frequently occur that a CNV is present (for example) in 3/200 cases and 0/200 controls. Such results are expected to occur by chance in a genome-wide search, and so do not necessarily imply a causal effect. (The observation of three independent, *de novo* structural mutations at the same locus in a disease cohort might be highly significant, because the rate of sporadic structural mutation seems to be much lower than the rate of CNV inheritance; such sporadic genomic disorders are discussed in an accompanying Perspective[2]).

It is natural to also consider the hypothesis that distinct CNVs at the same genomic locus may similarly influence disease risk in different individuals. An important precedent for such reasoning is the argument that diverse sequence variants in candidate genes are more frequently found in affected individuals than in controls[36,37]. In the case of rare coding SNPs, a framework is typically used in which nonsynonymous SNPs are examined based on their a priori likelihood of functionality. In the case of CNVs, similar paradigms may be useful: for example, pooling just those CNVs confirmed as affecting a candidate gene's coding sequence and nearby highly conserved elements. Although defining the right a priori criteria is not straightforward, the need for such criteria is: there is a great danger in (and long history of) *post hoc* explanations that can be invoked to support nonsignificant findings in discovery research.

Systematic biases can lead to false association

Years of SNP association studies — the vast majority of which proved irreproducible — have led to increased awareness of the factors that cause artifactual associations between genetic variants and phenotypes. CNP association studies are equally susceptible to these artifacts, which include population stratification, technical artifacts attributable to variability in the quality of DNA samples, and the general problem (inherent in all genome-wide studies) of distinguishing true signals from a genome-wide distribution of statistical sampling fluctuations.

Many phenotypes are associated with continental ancestry, and many CNPs (like many SNPs) vary in their frequencies across populations[24–26]. In disease association studies, such variants can be associated with disease owing to the confounding effect of ancestry (known as population stratification). Even in a study of individuals of European ancestry, variants that differ in frequency between northern and southern Europeans (such as the lactase persistence allele) can be artifactually associated with phenotypes (such as stature) that differ between northern and southern Europeans[38]. Methods to correct for stratification have been developed[39,40] and require the investigator to obtain extensive genetic information beyond the locus in question; it would seem reasonable to require such analyses in any CNP-based genome-wide association study, as in any SNP-based study. Family-based designs are another way to prevent stratification.

Disease association studies often utilize DNA that has been collected at a variety of clinical sites, extracted by different techniques, and prepared or assayed at different times. To the extent that DNA samples from affected individuals and controls differ, systematic technical bias can be introduced between the two groups. Some SNP assays are sensitive to DNA quality in ways that bias toward a particular result in lower-quality samples and can thereby lead to artifactual associations with disease[41]. This observation seems certain to apply to CNV studies as well. For example, the sensitivity of array comparative genomic hybridization (CGH) for detecting variants has been shown to vary from sample to sample based on variation in DNA and hybridization quality[29]. To the extent that altered copy numbers are undercalled in lower-quality DNA samples and hybridizations, heterogeneity in DNA preparation could lead to artifactual associations.

Although such biases may be sporadic and infrequent in focused, single-locus candidate-gene association studies, they are pervasive in genome-wide studies. This is because such studies involve looking for effects in the tails of a *P*-value distribution, where artifacts inevitably collect. As genome-wide CNP-disease association studies begin to be performed, it will be critical to seek out any systematic bias that distinguishes how DNA samples from affected individuals and controls are treated throughout the process of research. An important assessment is the extent to which the genome-wide distribution of *P* values conforms to the expected uniform distribution.

Perhaps the greatest cause of false association in the SNP literature has been the use of statistical thresholds inadequate to distinguish true associations from false positives. This is particularly problematic because of the low prior probability that any given variant (SNP or CNP) truly influences the trait of interest (at least, to an extent measurable with the sample size, technical approach and statistical framework employed)[42]. Just as in SNP association studies, it seems unlikely that an association of a CNP with disease that displays a *P* value of 0.05 will prove reproducible. On the other hand, a robustly significant *P* value (given the lower prior probability intrinsic in a genome-wide search), perhaps combined with functional data, surely can result in a compelling finding. Insistence upon the highest standards for proof in the early days of a field will save the community the consternation of irreproducible findings muddying the literature.

Opportunities

The coming years are likely to be tremendously exciting, as initial observations of common human common copy-number variation mature into an understanding that is crisp in molecular detail, complete in knowledge of location and frequency, and conducive to discoveries in the pathogenesis and genetic epidemiology of human disease. Perhaps the greatest impediments to such a future would be not the discovery of too few CNP-disease associations in the next year or two, but an insufficient investment in a truly effective set of tools and databases for their study, coupled with overly enthusiastic (but not quite reproducible) early claims of association with disease. With the proper focus and standards, CNP research will yield important insights, elucidating not only human genetic variation, but biological pathways and the mechanisms of human disease.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Inoue, K. & Lupski, J.R. Molecular mechanisms for genomic disorders. *Annu. Rev. Genomics Hum. Genet.* **3**, 199–242 (2002).
2. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43–S47 (2007).
3. Padiath, Q.S. *et al.* Lamin B1 duplications cause autosomal dominant leukodystrophy. *Nat. Genet.* **38**, 1114–1123 (2006).
4. Le Marechal, C. *et al.* Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat. Genet.* **38**, 1372–1374 (2006).
5. Lee, J.A. & Lupski, J.R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52**, 103–121 (2006).

6. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
7. Aitman, T.J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
8. 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.* **80**, 1037–1054 (2007).
9. Fanciulli, M. *et al. FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).
10. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
11. de Bakker, P.I. *et al.* Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
12. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38**, 663–667 (2006).
13. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
14. Reich, D.E., Gabriel, S.B. & Altshuler, D. Quality and completeness of SNP databases. *Nat. Genet.* **33**, 457–458 (2003).
15. Hinds, D.A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
16. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
17. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
18. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
19. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
20. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
21. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
22. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38**, 82–85 (2006).
23. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
24. Locke, D.P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
25. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38**, 1413–1418 (2006).
26. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
27. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. & Scherer, S.W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).
28. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
29. Fiegler, H. *et al.* Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res.* **16**, 1566–1574 (2006).
30. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
31. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
32. Lieberfarb, M.E. *et al.* Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. *Cancer Res.* **63**, 4781–4785 (2003).
33. Zhao, X. *et al.* An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).
34. Garraway, L.A. *et al.* Integrative genomic analyses identify *MITF* as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
35. Zhao, X. *et al.* Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. *Cancer Res.* **65**, 5561–5570 (2005).
36. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
37. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr. & Hobbs, H.H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
38. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **37**, 868–872 (2005).
39. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
40. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
41. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
42. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).