Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies

Xavier Estivill*, Lluís Armengol

Abstract

Genome-wide association scans (GWASs) using single nucleotide polymorphisms (SNPs) have been completed successfully for several common disorders and have detected over 30 new associations. Considering the large sample sizes and genome-wide SNP coverage of the scans, one might have expected many of the common variants underpinning the genetic component of various disorders to have been identified by now. However, these studies have not evaluated the contribution of other forms of genetic variation, such as structural variation, mainly in the form of copy number variants (CNVs). Known CNVs account for over 15% of the assembled human genome sequence. Since CNVs are not easily tagged by SNPs, might have a wide range of copy number variability, and often fall in genomic regions not well covered by whole-genome arrays or not genotyped by the HapMap project, current GWASs have largely missed the contribution of CNVs to complex disorders. In fact, some CNVs have already been reported to show association with several complex disorders using candidate gene/region approaches, underpinning the importance of regions not investigated in current GWASs. This reveals the need for new generation arrays (some already in the market) and the use of tailored approaches to explore the full dimension of genome variability beyond the single nucleotide scale.

Introduction

A large number of studies describing GWASs has been published recently. Several old and new associations have been detected by genotyping large collections of samples with hundred thousands of markers. Proof of concept of GWASs has been demonstrated and new biological pathways are now on the priority list of several investigators trying to understand asthma, Crohn disease, and diabetes, among other disorders. However, for most diseases, the identified genomic regions explain only a small fraction of the familial aggregation. Although these studies have been focused on SNPs as the common resource to explore genetic variability, other types of markers exist, which likely exert important phenotypic effects on gene expression and function. In this review, we explore the contribution of CNVs to common human disorders and evaluate the caveats of SNP-based GWASs in covering regions of the genome that have a high degree of plasticity and that could play an important role in disease susceptibility.

What Have We Missed in Current Genome-Wide Association Studies?

SNPs are the markers that have been selected to do the trick of uncovering the genetic determinants of complex traits and common disorders. This choice was mainly based on their abundance (over 12 million SNPs), and their use was boosted by the technological development of tools for high-throughput analysis of these variants. The Human Genome Project, followed by the HapMap Project [1] (http://www.hapmap.org/), has provided the landmark for the development of high-density SNP arrays to explore the nucleotide variability of the human genome, using powerful analytical methods based on statistical genetics, population genetics, and epidemiology.

Current association studies for common disorders and complex traits, aim to detect linkage disequilibrium (LD) between SNPs that genetically mark a given region (tagSNPs) and the functional variants (either at the RNA or protein level) responsible for the phenotypes. Due to their abundance and variability, SNPs have been considered powerful markers to identify loci underlying phenotypic variation in genetic association studies. To provide common and robust tools for

* To whom correspondence should be addressed. E-mail: xavier.estivill@crg.es
disease-associated gene discovery, the HapMap Consortium has genotyped nearly 4 million SNPs from individuals of the main human populations. A subset of these SNPs, covering the genome at the physical and genetic levels, is included in the commercially available arrays.

The outcome of the first round of studies involving thousands of patients and controls, and several hundred thousand SNPs has recently been published. GWASs have been completed for more than a dozen common disorders (Table 1) and several new associations have been detected. The Wellcome Trust Case Control Consortium (WTCCC) has reported the largest genome-wide association study performed so far, involving 14,000 patients and 3,000 controls [2]. Together with the WTCCC study, other publications (Table 1) have reported new, and confirmed previously known, statistically compelling associations for common disorders, including type 1 diabetes [2,3], type 2 diabetes [2–7], obesity [2], coronary heart disease [2,8–10], breast [11–13] and prostate [14,15] cancer, rheumatoid arthritis [2], Crohn disease [2,16–18], celiac disease [19], asthma [20], age-related macular degeneration [21], restless leg syndrome [22], and multiple sclerosis [23].

Table 1. Associations Identified in GWASs for Common Disorders Using Genotyping Arrays

<table>
<thead>
<tr>
<th>Disease</th>
<th>Array</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 diabetes</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>Custom (13K)</td>
<td>[3]</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[6]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap500 BeadChips</td>
<td>[4]</td>
</tr>
<tr>
<td></td>
<td>Illumina Human1 and HumanHap300 BeadChips</td>
<td>[7]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[5]</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>Custom (100K)</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td>Custom (100K)</td>
<td>[10]</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Illumina HumanHap500 BeadChips</td>
<td>[12]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[13]</td>
</tr>
<tr>
<td></td>
<td>Perlegen Sciences (267K)</td>
<td>[11]</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap500 BeadChips</td>
<td>[15]</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td>Crohn disease</td>
<td>Affymetrix GeneChip 500K</td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[16]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[17]</td>
</tr>
<tr>
<td></td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[18]</td>
</tr>
<tr>
<td>Celiac disease</td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[19]</td>
</tr>
<tr>
<td>Asthma</td>
<td>Illumina HumanHap300 BeadChips</td>
<td>[20]</td>
</tr>
<tr>
<td>Age-related macular degeneration</td>
<td>Affymetrix GeneChip 100K</td>
<td>[21]</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>Affymetrix GeneChip 500K</td>
<td>[23]</td>
</tr>
<tr>
<td>Restless leg syndrome</td>
<td>Affymetrix GeneChip 500K</td>
<td>[22]</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pgen.0030190.t001
of variation. Segments “A” and larger, “B” vary in orientation in the inversion. Segments “C” and “D” show different types of variation.

regions surrounding the detected associations must be performed, and is already under way in some cases. This will help to detect variants with functional consequences and larger effects than those so far uncovered, even if they are rare in the population and account only for a subset of patients.

Third, it would be interesting to see if epistasis exists between functional variants once they have been detected. It is remarkable that the data obtained so far mainly show absence of epistasis between variants for the same disorder or groups of disorders. Specific screens should be performed to assess the additive nature of the genetic component of the identified associations.

Fourth, although the HapMap project has provided an excellent tool for genetic association studies, it is clear that the set of markers analyzed in GWASs do not cover the entire genome variability. Despite the large number of SNPs that have been selected to explore genetic association using LD measures [24], and the coverage of nearly 100% of the genome using between 0.5 and 1 million tagSNPs [25], some regions are likely to be missed. Certainly, there are regions not well covered in HapMap due to the lack of sequence information, and, in large part, to the presence of CNVs and segmental duplications [26,27]. This has caused commercial panels to be deficient in SNPs covering these regions. Thus, future studies trying to reveal a more complete set of genetic determinants will necessarily require a larger number of SNPs (many with low minor allele frequencies and covering “unsettled” regions) and even larger cohorts. It has been estimated that to identify the complete set of loci involved in the genetic susceptibility to common disorders, sample sizes in the range of 2,000 to 60,000, and denser genetic maps, over 1 million SNPs will be needed. Despite the claims for “denser and larger,” the relatively large sample size of the studies performed so far and the wide genome coverage achieved suggests that, for some of the most deeply investigated disorders, the common genetic variants that underpin their genetic component have already been identified.

It seems clear that some of these questions will be solved by simply analyzing larger sample sets with denser SNP arrays, and by resequencing loci showing associations in a large number of samples. However, it is obvious that we need to explore the genome for other sources of variability that could explain the strong genetic component of several of the common disorders. Among sources to be explored are noncoding RNAs, structural variants, and epigenetic changes.

Many Versions Account for the Human Genome Sequence

When the human genome sequence was publicized six years ago, it was openly claimed that genetic differences between individuals account for less than 0.1% of the DNA sequence [28,29], a total of about 3 million nucleotides. Certainly, the statement referred to, and inferred from, the types of markers that had been, until then, widely used to explore diversity, construct genetic maps, and identify the genes responsible for more than 2,000 human monogenic disorders (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM). These markers included “old and new” types of polymorphisms, comprising restriction fragment length polymorphisms [30], variable number of tandem repeats or minisatellites [31], short tandem repeats or microsatellites [32,33], insertion/deletion polymorphisms [34], and the over 12 million SNPs that have been deposited in the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP).

In the last three years, a new form of genetic variation has been extensively reported. Genome structural variation has been known at the cytogenetic and molecular levels for a long time [35–37], but its importance at a genome-wide scale was not discovered until recently [38,39], with the use of array-based comparative genomic hybridization and other types of genome-scanning technologies. This variability entails large segments of DNA, typically over one kilobase (kb) and up to several megabases (Mb) and it comprises insertions, deletions, translocations, and inversions of genomic material (Figure 1). So far, the most commonly identified types of variants are gains and losses of DNA, which are called CNVs [40]. Inversions are also likely to be important changes, with direct potential positional effects and suppression of meiotic recombination, but, with some exceptions [41,42], most efforts toward characterization of variants have so far been focused on other types of changes. Obviously, structural variants are not exclusive of humans and they have also been identified in other organisms [43,44].

Fifteen comprehensive studies have explored structural variation in the human genome [38–41,45–52] (Table 2). These studies have used several approaches, mainly bacterial artificial chromosome (BAC) arrays, oligonucleotide arrays, SNP arrays, genotyping data, and computational alignment of genome sequences. There is wide variation of the coverage provided by the different methods and the level of polymorphism detected in the different studies (Figure 2). Many reasons account for these differences, including type of platform, genomic coverage, source of DNA samples (cell lines or fresh samples), control samples used by the different
projects, algorithms employed, and statistical thresholds. Comparison of experimental platforms, algorithms, and published surveys has recently been reviewed [53]. It is clear that the analysis of structural variants is still in its infancy, as compared to SNPs, but we have to admit that CNV analyses have additional complexity, due to their heterogeneity and the poor coverage that they exhibit in the assembled individual genomes [47].

The compilation of all reported variable regions is provided at several Web sites, including the UCSC (http://genome.ucsc.edu) and Ensembl (http://www.ensembl.org/) genome browsers, and the most updated summary can be found at the Database of Genomic Variants (http://projects.tcag.ca/variation/), which lists 8,083 CNVs that correspond to 3,933 loci in the human genome assembly (7 September 2007).

After the initial discovery that CNVs are common in the population, it was envisioned that CNVs might be traced using SNPs as proxies for different alleles of the structural changes. Although this is the case for some simple biallelic CNVs [40], the most common and polymorphic ones have a complex inheritance pattern and the SNPs located within do not always show Mendelian inheritance or are not in Hardy–Weinberg equilibrium. As a result of this, and also because of their identity with related sequences due to segmental duplications, many SNPs located at CNVs do not fulfill quality-control criteria and have been discarded in the design or in the analysis of genotyping experiments. Non-Mendelian behavior has also posed difficulties in the use of SNPs for tagging the inheritance of such variants. However, this abnormal behavior of markers has been used to successfully identify polymorphic deletions and inversions [42,50–52].

Since CNVs are not tagged easily by SNPs, many fall in regions that are not well defined in the available human genome sequence, and SNP content in commercial platforms is skewed towards “genotypable” SNPs present in the HapMap, it is likely that most GWASs (Table 3) have missed the potential contribution of CNVs to complex disorders. As mentioned above, our understanding of the organization of CNVs and their heritability is still very rudimentary. CNVs are likely to affect recombination, and the relationship with other markers might be relevant for common CNVs.

The current knowledge of CNVs is far from complete, because technological limitations of the approaches used so far to ascertain them have introduced an important bias towards medium-to-large-size CNVs. While technology has done very well for CNVs of sizes above 50 kb, smaller CNVs have hardly been detected. As further studies are performed covering regions below the 50-kb range, it is expected that a large number of additional CNVs, likely on the order of tens of thousands, will be detected (Figure 3). Considering the current human genome assembly, structural variants cover about 15% of the sequence (over 500 Mb). This figure is, however, imprecise, due to the lack of consensus in boundaries of CNV regions, the low level of resolution of clone arrays, and the near absence of replication of the reported data. On the basis of the expected size distribution of CNVs, they could likely affect up to one gigabase of sequence (≈1/3 of the genome). What is clear so far is that there is not a single human genome sequence and that several configurations, with alternative sequences at CNV regions, are present in the human population. Technologies that are able to screen the genome below this resolution will be essential. This should involve arrays specifically designed to interrogate at the 1–50-kb scale and sequencing specific regions with methods that allow the selection of DNA without previous knowledge of the sequence. In addition, efforts towards sequencing the genomes of different individuals to uncover their variability at the structural level are under way [54].

**Rare and Common CNVs Are Involved in Complex Disorders**

CNVs have already been shown to be associated with several complex/common disorders. Interestingly, most of these findings have been obtained by specific analysis of candidate genes or regions. Rare CNVs have been detected in some families of patients affected by Parkinson disease, Alzheimer disease, and chronic pancreatitis. Multiple cases of

---

**Table 2. Summary of Genome Scans to Study Structural Variations and CNVs of the Human Genome**

<table>
<thead>
<tr>
<th>Study</th>
<th>Method—Coverage</th>
<th>Detects</th>
<th>Samples</th>
<th>Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>[39]</td>
<td>ROMA—50,000 oligonucleotides</td>
<td>Deletion/insertion</td>
<td>20</td>
<td>76</td>
</tr>
<tr>
<td>[40]</td>
<td>Affymetrix 500K GeneChip</td>
<td>Deletion/insertion</td>
<td>270</td>
<td>1,203*</td>
</tr>
<tr>
<td>[41]</td>
<td>Illumina Human1 and HumanHap300 BeadChip</td>
<td>Deletion/insertion</td>
<td>182</td>
<td>340</td>
</tr>
<tr>
<td>[43]</td>
<td>Computational alignment between genomes</td>
<td>Deletion/insertion</td>
<td>180</td>
<td>586</td>
</tr>
<tr>
<td>[44]</td>
<td>Computational alignment between genomes</td>
<td>Deletion/insertion</td>
<td>269</td>
<td>541</td>
</tr>
<tr>
<td>[45]</td>
<td>BAC array—5,000 clones</td>
<td>Deletion/insertion</td>
<td>55</td>
<td>255</td>
</tr>
<tr>
<td>[46]</td>
<td>BAC array—2,000 clones</td>
<td>Deletion/insertion</td>
<td>47</td>
<td>160</td>
</tr>
<tr>
<td>[47]</td>
<td>Computational alignment between genomes</td>
<td>Deletion/insertion</td>
<td>95</td>
<td>3,654</td>
</tr>
<tr>
<td>[48]</td>
<td>BAC array—26,000 clones</td>
<td>Deletion/insertion</td>
<td>269</td>
<td>222</td>
</tr>
<tr>
<td>[49]</td>
<td>BAC array—2,000 clones</td>
<td>Deletion/insertion</td>
<td>270</td>
<td>1,116*</td>
</tr>
<tr>
<td>[50]</td>
<td>BAC array—5,000 clones</td>
<td>Deletion/insertion</td>
<td>180</td>
<td>586</td>
</tr>
<tr>
<td>[51]</td>
<td>Computational alignment between genomes</td>
<td>Deletion/insertion</td>
<td>269</td>
<td>541</td>
</tr>
<tr>
<td>[52]</td>
<td>Computational alignment between genomes</td>
<td>Deletion/insertion</td>
<td>269</td>
<td>176</td>
</tr>
<tr>
<td>[53]</td>
<td>BAC array—26,000 clones</td>
<td>Deletion/insertion</td>
<td>55</td>
<td>255</td>
</tr>
<tr>
<td>[54]</td>
<td>BAC array—2,000 clones</td>
<td>Deletion/insertion</td>
<td>55</td>
<td>255</td>
</tr>
</tbody>
</table>

* Correspond to copy number variant regions (CNVRs), each involving several CNVs. Detailed information on CNVs and references can be found at http://projects.tcag.ca/variation/. doi:10.1371/journal.pgen.0030190.t002
patients with Parkinson disease due to genomic duplication or triplication of the alpha-synuclein gene (SNCA) have been reported to cause hereditary early-onset parkinsonism with dementia, demonstrating a direct relationship between SNCA gene dosage and disease progression [55–57]. Similarly, several cases of duplication of the amyloid precursor protein (APP), with a role in familial Alzheimer disease and in Down syndrome brain neurodegeneration, have been described in families with early-onset Alzheimer dementia with cerebral amyloid angiopathy [58–60]. Finally, some members of families affected by hereditary pancreatitis have duplications or triplications of the cationic trypsinogen gene (PRSS) [61].

It is clear that in these three common disorders, the CNVs associated with the respective diseases represent rare events, and are not the major mechanism for disease susceptibility. Thus, rare genomic rearrangement events could affect common disorders in a manner similar to what has been reported for monogenic diseases, such as Neurofibromatosis type 1, for which large deletions are detected in about 10% of patients [62]. However, since rare CNVs are abundant in the genome, they could represent an important source of variability with which to explore the relationship between candidate genes and disease, and therefore to define new pathophysiology pathways.

Common CNVs have also been detected in people affected by certain other disorders. For example, variability in the susceptibility to HIV-1 infection has been related to copy number of the CCL3L1 gene [63]. Individuals with low copy numbers of the chemokine gene, relative to their ethnic background, are associated with markedly enhanced HIV-1/AIDS susceptibility. More recently, differences in copy number of the CCL3L1 chemokine have also been reported as a susceptibility factor for rheumatoid arthritis [64]. This region was not targeted in HapMap phases I and II and is not well covered by the Affymetrix and Illumina arrays; consequently, any attempt to perform association studies for HIV-1 susceptibility will likely fail in detecting a putative link with CCL3L1 copy variability (Figure 4). This region shows a large variability, not only in CCL3L1 copy number, but also in the genomic structure of individuals from different populations, as detected in the HapMap samples that have been genotyped [40]. In particular, the region is highly

Figure 2. Approaches Used for the Identification of CNVs and Other Types of Structural Changes in the Human Genome
Myriad methods and technologies have been employed to identify structural variants in the human genome. They are based on completely different experimental procedures and provide very different levels of resolution. The majority of findings (>80%) are attributable to a restricted number of high-throughput experiments with a limited resolution.

doi:10.1371/journal.pgen.0030190.g002
variable in the African population, with a large number of copy number gains in Africans and Asians, compared to Europeans.

Similarly, a copy number polymorphism including FCGR3 leads to a predisposition to glomerulonephritis in rats and humans, and to several types of autoimmune disorders, such as systemic lupus erythematosus (SLE), microscopic polyangiitis, and Wegener granulomatosis [65,66]. This region contains a complex 82-kb segmental duplication in the assembled genome sequence and CNVs have been detected in several studies in samples from the general population [40,46,48]. The coverage of the region is only partial in commercial arrays and the region of the CNVs and segmental duplication has a very low LD, with no blocks detected in HapMap populations.

Recently, another CNV region has been shown to be associated with SLE. Variable copy number of the complement component C4 (C4A and C4B) leads to different effects.

Table 3. Summary of Common Disorders for Which Associations to CNVs Have Been Reported

<table>
<thead>
<tr>
<th>Disorder</th>
<th>CNV</th>
<th>SD</th>
<th>SNPsa</th>
<th>Gene</th>
<th>Effect</th>
<th>Risk Associated</th>
<th>Study Type</th>
<th>Significance</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV-1/AIDS susceptibility</td>
<td>Common</td>
<td>Yes</td>
<td>No</td>
<td>CCL3L1</td>
<td>Dosage</td>
<td>Low CNV</td>
<td>Case control</td>
<td>Varies in populations</td>
<td>[63]</td>
</tr>
<tr>
<td>Rheumatoid arthritis and Type 1 diabetes</td>
<td>Common</td>
<td>Yes</td>
<td>No</td>
<td>CCL3L1</td>
<td>Dosage</td>
<td>High CNV</td>
<td>Case control</td>
<td>OR = 1.34; p = 0.009</td>
<td>[64]</td>
</tr>
<tr>
<td>SLE, microscopic Polyangiitis, and Wegener granulomatosis</td>
<td>Common</td>
<td>Yes</td>
<td>No</td>
<td>FCGR3B</td>
<td>Dosage</td>
<td>Low CNV</td>
<td>Case control</td>
<td>p &lt; 0.001</td>
<td>[65,66]</td>
</tr>
<tr>
<td>SLE</td>
<td>Common</td>
<td>Yes</td>
<td>No</td>
<td>C4A/C4B</td>
<td>Dosage</td>
<td>Low CNV</td>
<td>Case control</td>
<td>OR = 6.5; p &lt; 0.00002</td>
<td>[67]</td>
</tr>
<tr>
<td>Crohn disease</td>
<td>Common</td>
<td>Yes</td>
<td>No</td>
<td>DEFB4</td>
<td>Dosage</td>
<td>Low CNV</td>
<td>Case control</td>
<td>OR = 3.6; p &lt; 0.008</td>
<td>[71]</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>Common</td>
<td>No</td>
<td>Poor</td>
<td>GSK3B</td>
<td>Positional</td>
<td>High copy number</td>
<td>Case control</td>
<td>p = 0.002</td>
<td>[72]</td>
</tr>
<tr>
<td>Early-onset Parkinson disease</td>
<td>Rare</td>
<td>No</td>
<td>Yes</td>
<td>SNCA</td>
<td>Dosage</td>
<td>Duplication/triplication</td>
<td>Familial</td>
<td>NA</td>
<td>[55,56,57]</td>
</tr>
<tr>
<td>Hereditary early-onset Alzheimer disease</td>
<td>Rare</td>
<td>No</td>
<td>Yes</td>
<td>APP</td>
<td>Dosage</td>
<td>Duplication</td>
<td>Familial</td>
<td>NA</td>
<td>[59,60]</td>
</tr>
<tr>
<td>Hereditary pancreatitis</td>
<td>Rare</td>
<td>Yes</td>
<td>Poor</td>
<td>PRSS1</td>
<td>Dosage</td>
<td>Duplication</td>
<td>Familial</td>
<td>NA</td>
<td>[61]</td>
</tr>
<tr>
<td>Autism spectrum disorders</td>
<td>Common</td>
<td>NA</td>
<td>Vary</td>
<td>Multiple</td>
<td>Unknown</td>
<td>Higher “de novo” CNVs; multiple CNVs</td>
<td>Familial</td>
<td>NA</td>
<td>[78,79]</td>
</tr>
<tr>
<td>Familial breast cancer</td>
<td>Common</td>
<td>NA</td>
<td>Vary</td>
<td>MTUS5</td>
<td>Positional</td>
<td>Exon deletion confers lower risk</td>
<td>Familial</td>
<td>OR = 0.41; p &lt; 0.003</td>
<td>[73]</td>
</tr>
</tbody>
</table>

aSNP coverage in the Affymetrix and Illumina panels used in GWASs and HapMap genotyped SNPs. No, absence of SNPs; Yes, presence of SNPs; Poor, partial coverage; Vary, several regions with different types of coverage.

SD, segmental duplication; NA, not applicable; OR, odds ratio; p, p-value;
doi:10.1371/journal.pgen.0030190.g003

doi:10.1371/journal.pgen.0030190.g003

Figure 3. Expected and Observed Size Distribution of CNV Changes Identified to Date
Blue bars represent the frequencies of the currently identified CNVs in the size ranges depicted in the x-axis. A plausible scenario of variation in CNV size frequency is depicted as red vertical bars. An under-detection of variable fragments of small size (<50 kb) can be observed, which is likely due to technological limitations in the high-throughput assays used so far to identify CNVs, largely based on array CGH (Figure 2). Observed and expected CNVs that are >50 kb coincide, due to the powerful array methods, which cover the medium-to-large-size CNVs well. Dark blue bars represent the small-sized CNVs, which are more of a challenge to detect.
susceptibilities to SLE [67,68]. C4 gene copy number varies from two to six for total C4, zero to five for C4A, and zero to four for C4B. Compared with healthy subjects, patients with SLE clearly have lower copy numbers of C4 and C4A, and SLE susceptibility is significantly increased among subjects with only two copies of total C4 but decreased in those with more than five copies of C4 [68]. Interestingly, variability in copy number for the C4 genes and the genetic association to markers in this MHC region on Chromosome 6p21.32 has been known for several years [67,69,70], but their complex organization and their relationship with SLE has not yet been examined in detail. The C4A gene is fully contained in a 33-kb segmental duplication that shows 99.6% identity between copies in the assembled sequence of the human genome. The region has also been reported to be polymorphic in two studies exploring CNV regions [40,41]. This 80-kb region is not covered by the Affymetrix and Illumina arrays, and only three SNPs have been genotyped in HapMap, precluding positive association findings to these genes in whole-genome association studies (Figure 5).

Another report has detected variability in copy number of the beta defensin 2 gene (DEFB4) on Chromosome 8p23.1 in Crohn disease [71]. DEFB4 dosage is lower in colonic Crohn disease compared with controls, showing that a lower DEFB4 gene copy number predisposes to colonic Crohn disease through diminished beta-defensin expression. Again, for this locus, there is a cluster of segmental duplications, and most CNV studies have detected this region as being variable. This region, which spans about 1 Mb and contains a gap in the assembled genome sequence, has only four SNPs in the Affymetrix array and one in the Illumina array (not shown). Although the region was not detected in the GWAS for Crohn disease [2,16–18], it is obvious that this region was not satisfactorily covered by these arrays. Only the targeted analysis of the region using quantitative methods was able to uncover the link with Crohn disease [71].

Finally, several other studies exploring CNVs in common...
disorders are being performed and some findings in bipolar disorder [72] and breast cancer [73] have already been reported. Therefore, we expect that there will be a plethora of reports describing new associations between CNVs and common disorders and complex traits in the coming months to years.

A common feature of the regions for complex/common disorders identified so far is the presence of both CNVs and segmental duplications. A clear association between duplicons and CNVs in the human genome has been reported [40]. This association is stronger for CNVs that are multiallelic or have a complex pattern. Interestingly, all CNV loci that have been found associated with common disorders are both complex and multiallelic. Thus, the development of assays for common/common CNV loci could provide good tools for the analysis of common disorders.

The mechanisms by which CNVs could contribute to disease are numerous [74]. Due to their location and nature, a significant fraction of CNVs are likely to have functional consequences, either by gene dosage alteration, disruption of genes, positional effects, uncovering deleterious alleles, or modulating the action of other sequences. We still have limited evidence of the role of CNVs in gene expression. Stranger and colleagues [75] have examined RNA levels in lymphoblastoid cell lines from 210 unrelated HapMap individuals and have used CNV data from these samples generated by the Structural Variation Consortium [40] to conclude that 18% of the variation in expression levels of ∼15,000 genes is attributable to copy number differences. This study represents the first attempt to evaluate the genome-wide impact of SNPs and CNVs on gene expression. A potential explanation for the relatively low contribution of CNVs to variability of gene expression as compared to SNPs in the study of Stranger and colleagues [75] is the limited resolution of the arrays used and the wide definition of CNV regions considered in the analysis.

**Combination of SNP and CNV Genotyping in Common Disorders**

Although a large number of SNPs for regions containing CNVs are listed in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), most of them lack genotyping frequencies, have not been confirmed by other investigators, or fail during the design of multiplex genotyping assays. Many of these SNPs are located in segmental duplications and they correspond to paralogous sequence variants or SNPs that are copy specific [76]. As a consequence, most of these regions have systematically been excluded from the current high-throughput SNP typing assays.

Many investigators in the field of the genetics of common disorders have realized the need to cover other types of variants in their genome scans. Commercial genotyping companies (mainly Affymetrix and Illumina) are redesigning their platforms to allocate probes for CNV regions and they now claim a genome-wide coverage of known and new CNVs. While this reflects the recent attention that CNVs have attracted in the genotyping field, the reliability of the coverage and the capacity of these arrays to discriminate between a wide range of copies of a given CNV has yet to be proven. This discrimination capacity is one of the main challenges to extracting the complexity of genomic structural variability and will be crucial for association studies. On the other hand, companies dedicated to array-comparative genomic hybridization (CGH) production (Nimblegen and Agilent) are developing denser arrays that could explore the complete genome, also offering flexibility in the incorporation of probes for targeted studies. A review about the different platforms available for CNV analysis has recently been published [77]. There are many reasons for and against the use of one over another. While genotyping platforms provide two products for the price of one, CGH arrays provide better signal accuracy, because they compare real samples in the same experiment. The choice depends on the specific status of the project, especially if a GWAS has already been performed with first-generation genotyping arrays, which have poor coverage in CNV regions. In these cases, CGH arrays should provide coverage of CNVs missed by the genotyping platforms. Indeed, several efforts are under way to screen, using CGH platforms, the WTCCC samples already genotyped with Affymetrix arrays (Figure 6).

It is important to note that all the associations between CNVs and complex disorders reported so far have been unveiled through candidate gene or candidate region approaches. Indeed, only thorough investigations by groups working on the disorders or with specific interest in a concrete variable region have been able to dissect the fine spectrum of variability to provide a link with the phenotypes (Table 3). Although genotyping scans could be able to detect CNV regions, current approaches do not provide any kind of discrimination of the variability spectrum associated to these loci, and are therefore unable to distinguish copy numbers with respect to phenotype. Several methods allow quantification of CNVs, including multiplex ligation-dependent probe amplification (MLPA), multiplex amplification and probe hybridization (MAPH), quantitative multiplex PCR of short fluorescent fragment (QMPSF), dynamic allele-specific hybridization, semiquantitative fluorescence in situ hybridization (SQ-FISH), paralogue ratio test, and multiple amplicon quantification, among others (Figure 6). Precise definition of breakpoints can be achieved by PFGE (pulsed field gel electrophoresis), regular Southern blotting, and sequencing. Ultrasquencing technologies (based on synthesis, GS-FLEX [Roche-454] and 1G Solexa Genetic Analyzer [Solexa-Illimina]; or on ligation, SOLID [Applied Biosystems]) should also provide this level of resolution, but specific experimental trials have to be developed to achieve a successful resequencing and assembly.
A

B

Oligonucleotide Genotyping  Sequencing and Computation  Array-Comparative Genomic Hybridization

Southern blotting  PFGE  Fiber-FISH  FISH

qPCR  DASH  PRT  MLPA/MAPH  MAQ  QMPSF  Sequencing

doi:10.1371/journal.pgen.0030190.g006
of regions with the high level of plasticity and identity of CNVs. Tailored approaches to detect the variability in copy number of common CNV loci, and the use of genetic approaches that explore the differences between phenotypes at a whole-genome scale should be pursued. A diagram of genome-wide and locus-specific approaches to detect and analyze CNVs is proposed in Figure 6. Improvements in the field of CNVs are clearly needed both for the genome-wide coverage and for the precise quantification of specific CNVs.

Progress in the identification of CNVs associated with complex disorders will likely take place at a rapid pace in the next few months to years. Currently available tools will only be able to disclose variants that, because of their genomic (large rearrangements) and genetic characteristics (de novo cases), are easily discovered [78,79]. Thus, the systematic exploration of multiallelic CNVs, with precise characterization of copy numbers, should become essential when exploring the role of CNV in many traits and diseases. Finally, since many CNVs contain genes with an important role in adaptation to the environment and response to external effects [40], it is tempting to speculate that CNV alleles could have a major role in disease predisposition and response to drugs.

Conclusions

Recent progress in the identification of loci showing association to complex disorders has provided not only a proof of concept of GWASs, but has also led to the identification of several new biological associations. The need for larger sample sets and better coverage of genome variability at the nucleotide level, including resequencing, is likely to be achieved after this initial first round of GWASs. However, the complete spectrum of genomic variability will not be elucidated by this approach. Several CNVs have been shown to be implicated in common disorders, as rare and common genomic changes, providing biological support to several pathophysiological pathways. New types of arrays, covering CNVs and segmental duplications, will facilitate the identification of regions that contain CNVs, but will likely still fail to detect associations with a wide range of variability in copy number. A comprehensive tailored analysis of common and rare CNVs will not only complement GWASs using SNPs and sequencing, but will also provide a new, more powerful tool for examining the genetic components of common disorders and complex traits in humans and other organisms.

Acknowledgments

We thank Mónica Bayés and Mario Cáceres for critical reading of the manuscript.

Author contributions. XE and LA conceived the study, analyzed the data, and wrote the paper.

Funding. The laboratory of XE is supported by the Departament d'Educació i Universitats and the Departament de Salut of the Catalan Autonomous Government (Generalitat de Catalunya); the Ministry of Health and the Ministry of Education and Science of the Spanish government; the European Union Sixth Framework Programme; and Genoma España. LA is funded by the European Union AzETHiP Project (037627).

Competing interests. The authors have declared that no competing interests exist.

References


Figure 6. CNV Characterization Strategies

(A) Scales of resolution at the nucleotide level and maximum number of loci interrogated by the different methods (only the most widely used approaches are shown).

(B) Diagram of different approaches in CNV analysis, either at the genome-wide scale or at individual/multiplex loci. Arrows indicate the deeper analysis that is needed after initial detection by one methodology or another.

DASH, dynamic allele-specific hybridization [80]; PRT, paralogue ratio test [81]; MAQ, multiple amplicon quantification [82]; qPCR, quantitative PCR.


