OPEN ACCESS Freely available online

PLoS GENETICS

Review

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

Editor: Elizabeth M. C. Fisher, University College London, United Kingdom

Citation: Estivill X, Armengol L (2007) Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet 3(10): e190. doi:10.1371/journal.pgen.0030190

Abbreviations: BAC, bacterial artificial chromosome; CGH, comparative genomic hybridization; CNV, copy number variant; FISH, fluorescence in situ hybridization; GWAS, genome-wide association scan; LD, linkage disequilibrium; MAPH, multiplex amplification and probe hybridization; MLPA, multiplex ligation-dependent probe amplification; PFGE, pulsed field gel electrophoresis; QMPSF, quantitative multiplex PCR of short fluorescent fragment; ROMA, representational oligonucleotide microarray analysis; SLE, systemic lupus erythematosus; SNP, single nucleotide polymorphism; SQ-FISH, semiquantitative fluorescence in situ hybridization; WTCCC, Wellcome Trust Case Control Consortium

Xavier Estivill and Lluís Armengol are with the Genes and Disease Program, Center for Genomic Regulation (CRG), National Genotyping Center (CeGen), CIBERESP, and Pompeu Fabra University (UPF), Charles Darwin Square Parc de Recerca Biomedica Building (PRBB), Barcelona, Catalonia, Spain.

* To whom correspondence should be addressed. E-mail: xavier.estivill@crg.es

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

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

doi:10.1371/journal.pgen.0030190.t001

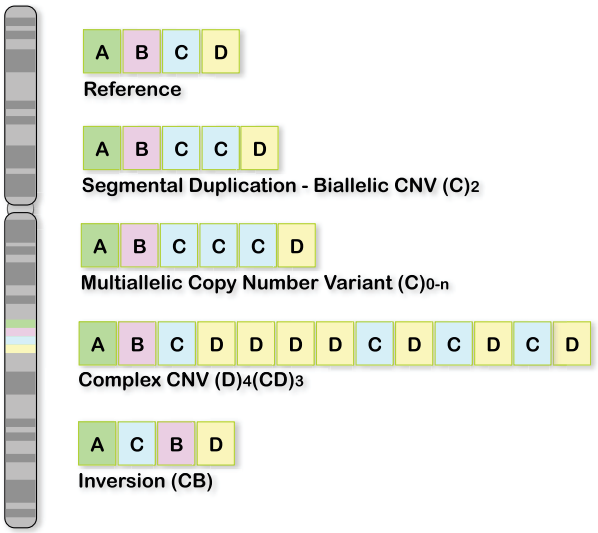
disease-associated gene discovery, the HapMap Consortium has genotyped nearly 4 million SNPs from individuals of the main human populations. A subset of these SNPs, covering the genome at the physical and genetic levels, is included in the commercially available arrays.

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, for seven diseases 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 [2,3] and 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].

The above-mentioned analyses represent an obvious step forward in the arena of the study of the genetic contribution to complex diseases and have undoubtedly proved the utility of the GWAS approach using SNPs to identify new genetic associations without previous hypotheses about their biology. Each of these reports has described links with known or new biological pathways, and has also established novel mechanistic connections among pathways and among disorders. The set of loci reported so far should potentiality facilitate progress in the understanding of the physiology of each of these disorders. These studies, however, raise several questions in relation to the genetic basis of complex diseases and the strategies used so far towards the identification of a complete set of susceptibility loci.

First, it is obvious that the genetic picture obtained for each of these disorders, even for those targeted by independent cohorts, such as in the case of type 2 diabetes, is still far from complete. The identified associations, with some exceptions (the major histocompatibility complex, MHC, locus), have a modest effect with odds ratios lower than 1.5. Thus, the nine confirmed loci for type 2 diabetes [2–7] might explain about 3% of the genetic variance, and 14 loci identified for Crohn disease [2,16–18] cover less than 10% of the variance. If we take into account the outcome achieved in these studies using such large number of samples and SNPs, it is expected that new associations for common disorders using SNP markers will likely have similar or even lower effects, and association values will likely not go far above current figures. Furthermore, we are uncertain about how well the additional small effects will be able to disclose the strong heritability that many complex disorders exhibit.

Second, it is also obvious from the studies reported so far (with the exception of age-related macular degeneration [21] and some other disorders), that the identified variants are not the functional ones. Thus, the role of most genetic changes in the molecular basis of disorders has not yet been discovered. Sequencing of a large number of patients with the aforementioned disorders along with a deep coverage of the



doi:10.1371/journal.pgen.0030190.g001

Figure 1. Types of Genomic Structural Changes Affecting Segments of DNA, Leading to Deletions, Duplications, Inversions, and CNV Changes (Biallelic, Multillelic, and Complex)

The only segment that is constant is "A." Segment "B" varies 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 loses 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

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

Study	Method—Coverage	Detects	Samples	Variants
[38]	BAC array—5,000 clones	Deletion/insertion	55	255
[45]	BAC array—2,000 clones	Deletion/insertion	47	160
[46]	BAC array—2,000 clones	Deletion/insertion	269	222
[40]	BAC array—26,000 clones	Deletion/insertion	270	1,116[a]
[48]	BAC array—26,000 clones	Deletion/insertion	95	3,654
[40]	Affymetrix 500K GeneChip	Deletion/insertion	270	1,203[a]
[49]	Illumina Human1 and HumanHap300 BeadChip	Deletion/insertion	182	340
[39]	ROMA—85,000 oligonucleotides	Deletion/insertion	20	76
[50]	Genotyping data—100–200 Mb	Deletion	24	215
[52]	Genotyping data—1.3 million genotypes	Deletion	180	586
[51]	Genotyping data—1.3 million genotypes	Deletion	269	541
[42]	Genotyping data—1.3 million genotypes	Inversions	269	176
[41]	Fosmid end sequencing/mapping	Deletion/insertion/inversion	1	297
[34]	Computational alignment between genomes	Deletion/insertion/inversion	36	294,498
[47]	Computational alignment between genomes	Deletion/insertion/inversion	2	13,534

[a]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

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).

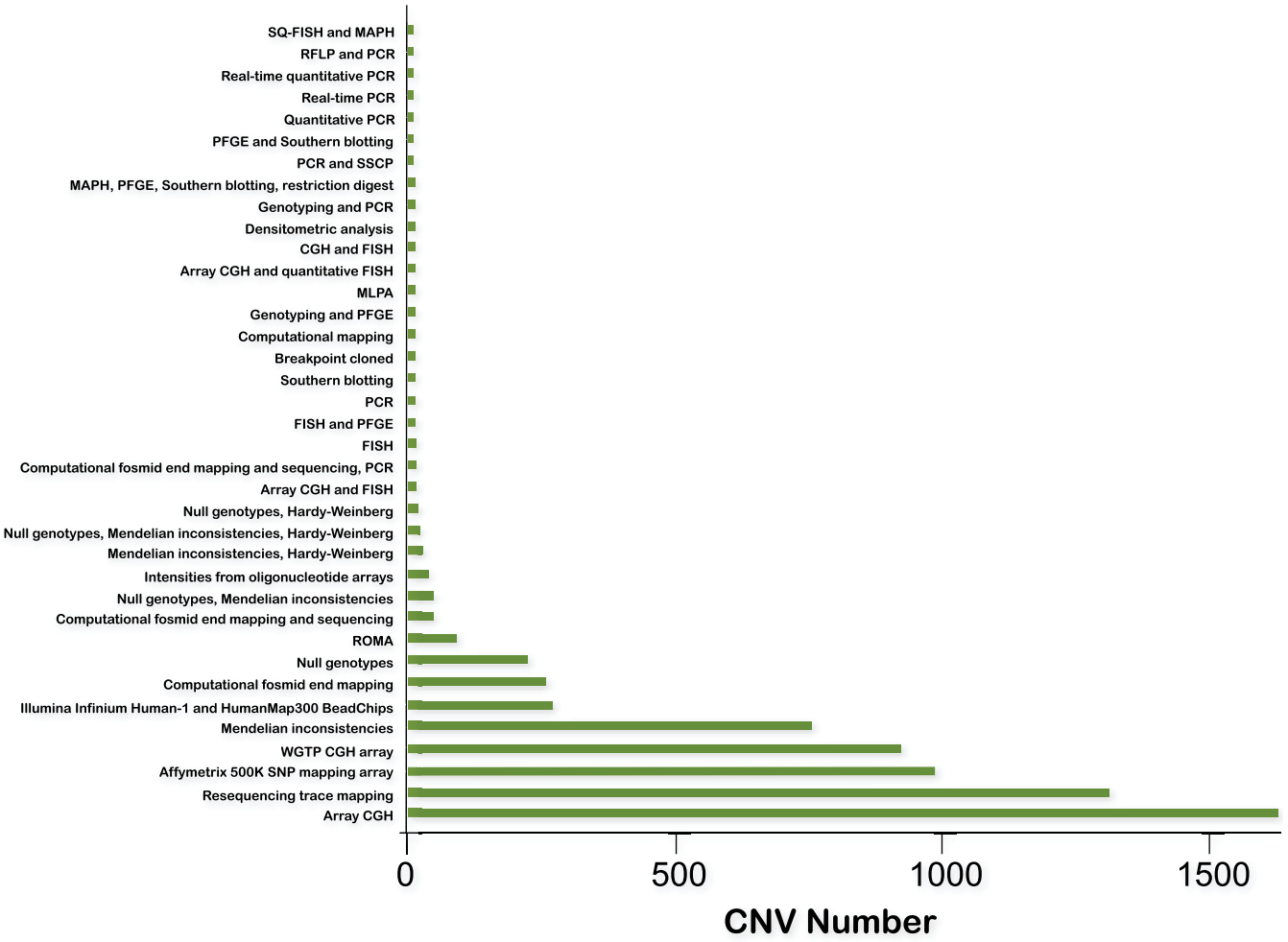
After the initial discovery that CNVs are common in the population, it was envisioned that CNVs might be traced using SNPs as proxies for different alleles of the structural changes. Although this is the case for some simple biallelic CNVs [40], the most common and polymorphic ones have a complex inheritance 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



doi:10.1371/journal.pgen.0030190.g002

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.

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

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

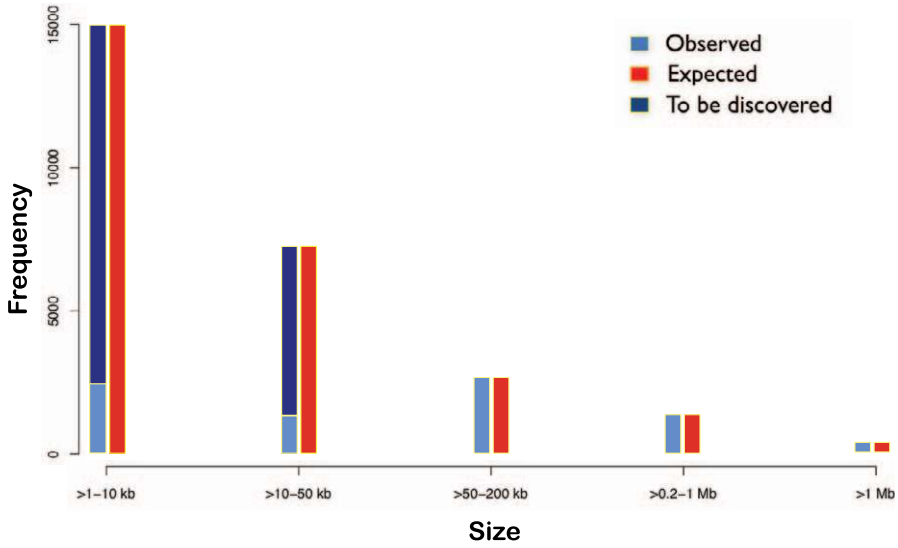
Disorder	CNV	SD	SNPs[a]	Gene	Effect	Risk Associated	Study Type	Significance	Reference
HIV-1/AIDS susceptibility	Common	Yes	No	*CCL3L1*	Dosage	Low CNV	Case control	Varies in populations	[63]
Rheumatoid arthritis and Type 1 diabetes	Common	Yes	No	*CCL3L1*	Dosage	High CNV	Case control	OR = 1.34; $p = 0.009$	[64]
SLE, microscopic Polyangiitis, and Wegener granulomatosis	Common	Yes	No	*FCGR3B*	Dosage	Low CNV	Case control	$p < 0.001$	[65,66]
SLE	Common	Yes	No	*C4A/C4B*	Dosage	Low CNV	Case control	OR = 6.5; $p < 0.00002$	[67]
Crohn disease	Common	Yes	No	*DEFB4*	Dosage	Low CNV	Case control	OR = 3.6 $p < 0.008$	[71]
Bipolar disorder	Common	No	Poor	*GSK3B*	Positional	High copy number	Case control	$p = 0.002$	[72]
Early-onset Parkinson disease	Rare	No	Yes	*SNCA*	Dosage	Duplication/triplication	Familial	NA	[55,56,57]
Hereditary early-onset Alzheimer disease	Rare	No	Yes	*APP*	Dosage	Duplication	Familial	NA	[59,60]
Hereditary pancreatitis	Rare	Yes	Poor	*PRSS1*	Dosage	Triplication	Familial	NA	[61]
Autism spectrum disorders	Common	NA	Vary	Multiple	Unknown	Higher "de novo" CNVs; multiple CNVs	Familial	NA	[78,79]
Familial breast cancer	Common	No	Yes	*MTUS1* (exon 4)	Positional	Exon deletion confers lower risk	Familial	OR = 0.41; $p < 0.003$	[73]

[a]SNP 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.t003

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.

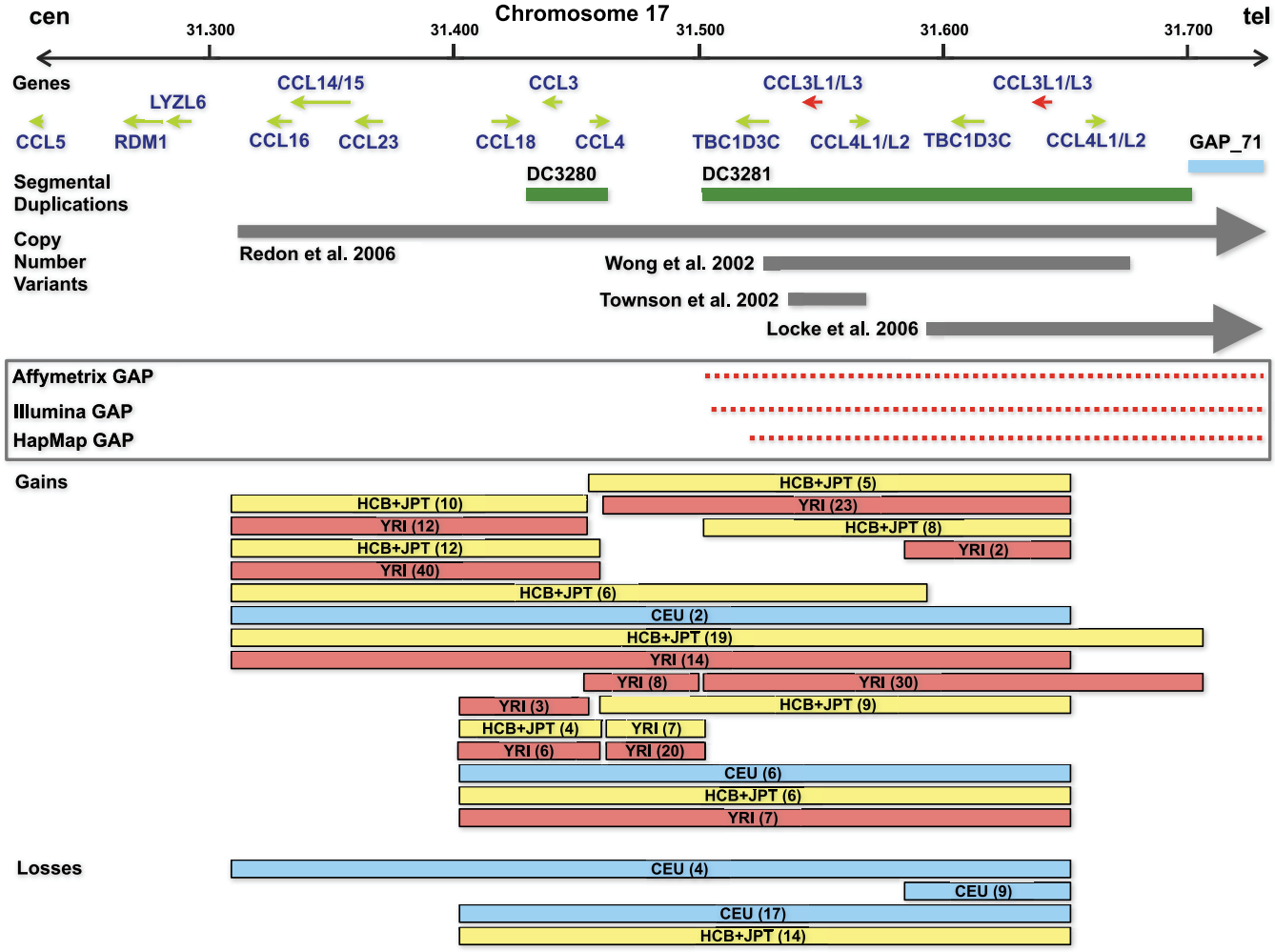
Recently, another CNV region has been shown to be associated with SLE. Variable copy number of the complement component *C4* (*C4A* and *C4B*) leads to different



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.



doi:10.1371/journal.pgen.0030190.g004

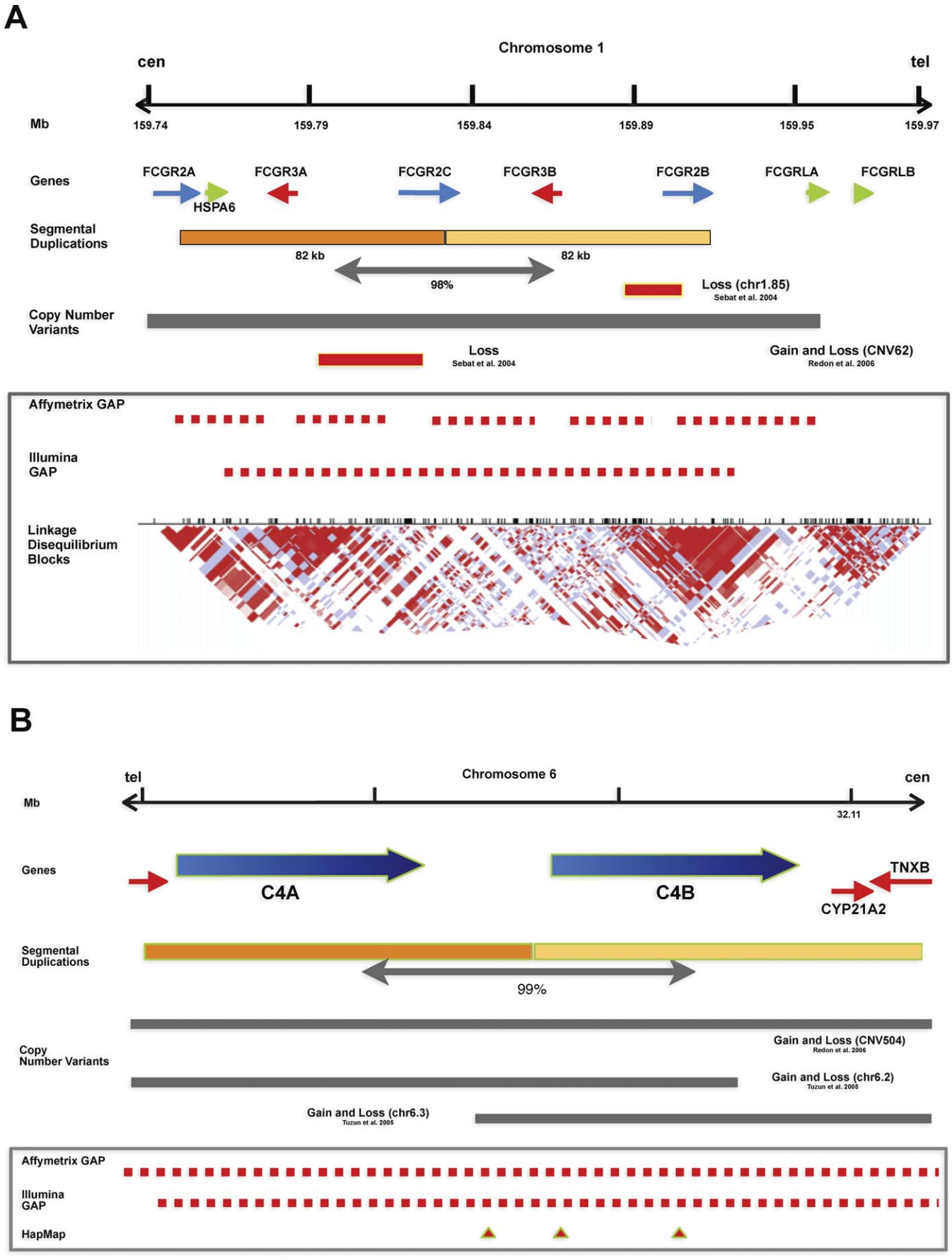
Figure 4. Genomic Organization of the Chemokine Cluster on Human Chromosome 17, Containing the *CCL3L1* Gene (Red Arrows), Which Shows Variability in Copy Number and Association to HIV-1 Infectivity and AIDS Susceptibility

This region contains several segmental duplications and has been reported to vary in copy number in several studies. The Affymetrix 500K and Illumina HumanHap 550 arrays do not cover this region well, and completely lack SNPs in the *CCL3L1/L3* gene (red dotted lines). A large number of gains and losses have been reported in the HapMap samples. Numbers in parentheses indicate the number of events involving genomic changes.
CEU, European; HCB, Chinese; JPT, Japanese; YRI, African.

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



doi:10.1371/journal.pgen.0030190.g005

Figure 5. Schematic Representation of Two Genomic Regions That Involve CNVs Associated with SLE [65,66]

(A) The region of Chromosome 1 containing the *FCGR3* gene cluster is highly variable and contains segmental duplications with a high sequence identity. Several CNVs have been reported that span this region. The genomic organization of the cluster is highly complex and not well solved in the current assembly of the genome sequence. The Affymetrix 500K and Illumina HumanHap 550 arrays do not cover this region well (red dotted lines). (B) The region of Chromosome 6p21, containing the *C4A* and *C4B* genes, is embedded in a region of complex genomic organization [67,69,70]. The region has been shown to contain segmental duplications and CNVs. The Affymetrix 500K and Illumina HumanHap 550 genotyping platforms do not cover this region, either (red dotted lines).

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/complex 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.

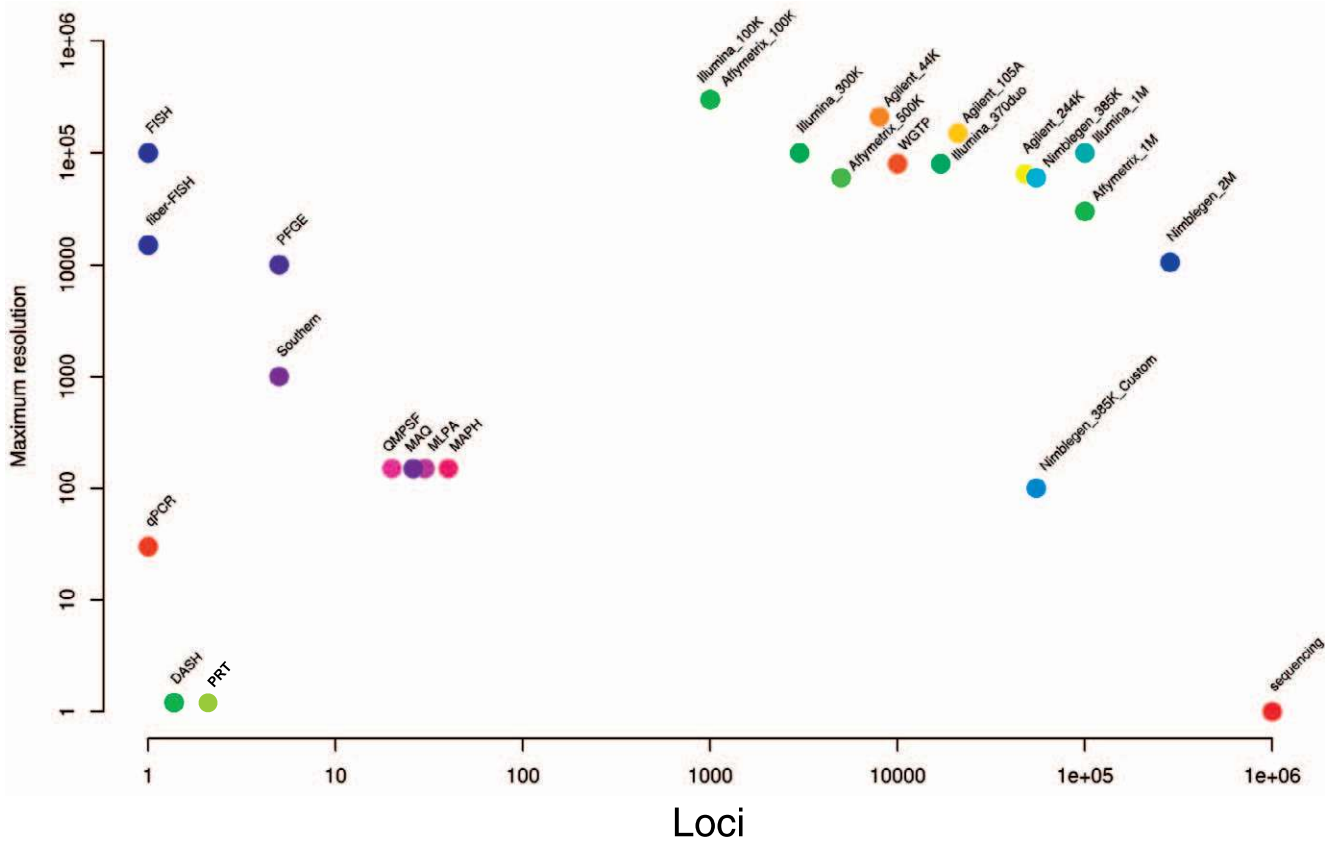
Combination of SNP and CNV Genotyping in Common Disorders

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.

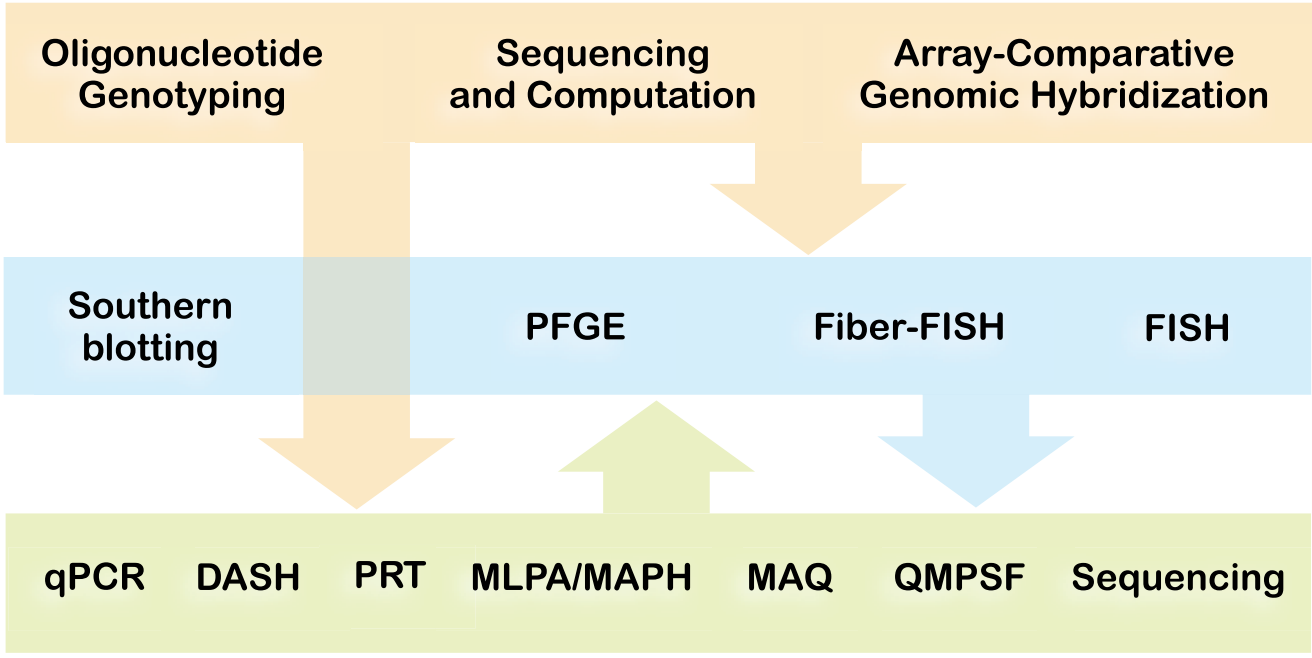
Many investigators in the field of the genetics of common disorders have realized the need to cover other types of variants in their genome scans. Commercial genotyping companies (mainly Affymetrix and Illumina) are redesigning their platforms to allocate probes for CNV regions and they now claim a genome-wide 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. Ultrasequencing technologies (based on synthesis, GS-FLEX [Roche-454] and 1G Solexa Genetic Analyzer [Solexa-Illumina]; 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



doi:10.1371/journal.pgen.0030190.g006

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.

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 Mònica 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 AnEUploidy Project (037627).

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

References

1. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
2. The Welcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
3. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, et al. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nat Genet 38: 617–619.
4. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331–1336.
5. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, et al. (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 39: 770–775.
6. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341–1345.
7. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, et al. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881–885.
8. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32: 650–654.
9. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, et al. (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316: 1491–1493.
10. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, et al. (2007) A common allele on chromosome 9 associated with coronary heart disease. Science 316: 1488–1491.
11. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087–1093.
12. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870–874.
13. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, et al. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39: 865–869.
14. Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, et al. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39: 631–637.
15. Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, et al. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39: 645–649.
16. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314: 1461–1463.
17. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, et al. (2007) Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet 3: e58.
18. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39: 596–604.
19. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39: 827–829.
20. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.
21. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
22. Winkelmann J, Schormair B, Lichtner P, Ripke S, Xiong L, et al. (2007) Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet 39: 1000–1006.

23. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med 357: 851–862.
24. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. Nat Genet 37: 1217–1223.
25. de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 38: 1298–1303.
26. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, et al. (2002) Recent segmental duplications in the human genome. Science 297: 1003–1007.
27. Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, et al. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4: R25.
28. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
29. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
30. Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci U S A 75: 5631–5635.
31. Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, et al. (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235: 1616–1622.
32. Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44: 397–401.
33. Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44: 388–396.
34. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, et al. (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16: 1182–1190.
35. Craig-Holmes AP, Moore FB, Shaw MW (1973) Polymorphism of human C-band heterochromatin. I. Frequency of variants. Am J Hum Genet 25: 181–192.
36. Estivill X, Farrall M, Scambler PJ, Bell GM, Hawley KM, et al. (1987) A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. Nature 326: 840–845.
37. Goossens M, Dozy AM, Embury SH, Zachariades Z, Hadjiminas MG, et al. (1980) Triplicated alpha-globin loci in humans. Proc Natl Acad Sci U S A 77: 518–521.
38. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004) Detection of large-scale variation in the human genome. Nat Genet 36: 949–951.
39. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
40. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
41. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727–732.
42. Bansal V, Bashir A, Bafna V (2007) Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res 17: 219–230.
43. Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, et al. (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 3: e3. doi:10.1371/journal.pgen.0030003
44. Li J, Jiang T, Mao JH, Balmain A, Peterson L, et al. (2004) Genomic segmental polymorphisms in inbred mouse strains. Nat Genet 36: 952–954.
45. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al. (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77: 78–88.
46. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79: 275–290.
47. Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, et al. (2006) Genome assembly comparison identifies structural variants in the human genome. Nat Genet 38: 1413–1418.
48. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al. (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80: 91–104.
49. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, et al. (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16: 1–14.
50. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38: 82–85.
51. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, et al. (2006) Common deletion polymorphisms in the human genome. Nat Genet 38: 86–92.
52. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 38: 1251–1260.
53. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, et al. (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39: S7–S15.
54. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. (2007) Completing the map of human genetic variation. Nature 447: 161–165.
55. Ibanez P, Bonnet AM, Debarges B, Lohmann E, Tison F, et al. (2004) Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. Lancet 364: 1169–1171.
56. Chartier-Harlin MC, Kachergus J, Roumier C, Mouroux V, Douay X, et al. (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. Lancet 364: 1167–1169.
57. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, et al. (2003) alpha-Synuclein locus triplication causes Parkinson's disease. Science 302: 841.
58. Sleegers K, Brouwers N, Gijselinck I, Theuns J, Goossens D, et al. (2006) APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. Brain 129: 2977–2983.
59. Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, et al. (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38: 24–26.
60. Cabrejo L, Guyant-Marechal L, Laquerriere A, Vercelletto M, De la Fourniere F, et al. (2006) Phenotype associated with APP duplication in five families. Brain 129: 2966–2976.
61. Le Marechal C, Masson E, Chen JM, Morel F, Ruszniewski P, et al. (2006) Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat Genet 38: 1372–1374.
62. Lopez Correa C, Brems H, Lazaro C, Estivill X, Clementi M, et al. (1999) Molecular studies in 20 submicroscopic neurofibromatosis type 1 gene deletions. Hum Mutat 14: 387–393.
63. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, et al. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307: 1434–1440.
64. McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, et al. (2007) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. Ann Rheum Dis. E-pub ahead of print. doi: 10.1136/ard.2007.075028
65. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, et al. (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature 439: 851–855.
66. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721–723.
67. Yang Y, Chung EK, Zhou B, Lhotta K, Hebert LA, et al. (2004) The intricate role of complement component C4 in human systemic lupus erythematosus. Curr Dir Autoimmun 7: 98–132.
68. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al. (2007) 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.
69. Chung EK, Yang Y, Rennebohm RM, Lokki ML, Higgins GC, et al. (2002) Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. Am J Hum Genet 71: 823–837.
70. Chung EK, Yang Y, Rupert KL, Jones KN, Rennebohm RM, et al. (2002) Determining the one, two, three, or four long and short loci of human complement C4 in a major histocompatibility complex haplotype encoding C4A or C4B proteins. Am J Hum Genet 71: 810–822.
71. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, et al. (2006) 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 79: 439–448.
72. Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, et al. (2007) Increase in GSK3beta gene copy number variation in bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 144: 259–265.
73. Frank B, Hemminki K, Meindl A, Wappenschmidt B, Sutter C, et al. (2007) BRIP1 (BACH1) variants and familial breast cancer risk: a case-control study. BMC Cancer 7: 83.
74. Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8: 639–646.
75. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.
76. Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, et al. (2002) Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum Mol Genet 11: 1987–1995.
77. Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39: S16–S21.
78. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445–449.

79. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, et al. (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet 39: 319–328.
80. Stromqvist Meuzelaar L, Hopkins K, Liebana E, Brookes AJ (2007) DNA diagnostics by surface-bound melt-curve reactions. J Mol Diagn 9: 30–41.
81. Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, et al. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res 35: e19. doi:10.1093/nar/gkl1089
82. Sutrala SR, Goossens D, Williams NM, Heyrman L, Adolfsson R, et al. (2007) Gene copy number variation in schizophrenia. Schizophr Res. doi:10.1016/j.schres.2007.07.029