Copy number variants, diseases and gene expression

Charlotte N. Henrichsen, Evelyne Chaignat and Alexandre Reymond*

The Center for Integrative Genomics, Genopode Building, University of Lausanne, CH-1015 Lausanne, Switzerland

Copy number variation (CNV) has recently gained considerable interest as a source of genetic variation likely to play a role in phenotypic diversity and evolution. Much effort has been put into the identification and mapping of regions that vary in copy number among seemingly normal individuals in humans and a number of model organisms, using bioinformatics or hybridization-based methods. These have allowed uncovering associations between copy number changes and complex diseases in whole-genome association studies, as well as identify new genomic disorders. At the genome-wide scale, however, the functional impact of CNV remains poorly studied. Here we review the current catalogs of CNVs, their association with diseases and how they link genotype and phenotype. We describe initial evidence which revealed that genes in CNV regions are expressed at lower and more variable levels than genes mapping elsewhere, and also that CNV not only affects the expression of genes varying in copy number, but also have a global influence on the transcriptome. Further studies are warranted for complete cataloguing and fine mapping of CNVs, as well as to elucidate the different mechanisms by which they influence gene expression.

INTRODUCTION

A fundamental question in current biomedical research is to establish a link between genomic variation and phenotypic differences, which encompasses both the seemingly neutral polymorphic variation, as well as the pathological variation that causes or predisposes to disease. DNA copy number variants, defined as stretches of DNA larger than 1 kb that display copy number differences in the normal population (1), have recently gained considerable interest as a source of genetic diversity likely to play a role in functional variation (2–5).

Extensive studies were performed to establish CNV catalogs for multiple species, however little is known about the functional impact of CNVs at the cellular and organismal level. Here, we review recent works on the identification of CNVs, their association with disease and their global impact on gene expression, as well as discuss possible mechanisms by which CNVs exert this influence.

COPY NUMBER VARIATION IN HUMAN AND MODEL ORGANISMS

Whole-genome CNV catalogs have been established for the human (4,6–8), the mouse (9–16), the rat (17), the chimpanzee (7,18), the rhesus macaque (19) and Drosophila melanogaster (20,21) (Table 1). CNVs were identified by array comparative genome hybridization (aCGH) (22), either with bacterial artificial chromosome or oligonucleotide probes. Alternatively, bioinformatics mining of whole genome shotgun (WGS) sequences was used to predict CNVs in human, macaque and rat (17,23,24). As observed in humans, distinct array platforms yielded considerably different datasets with little overlap [43% and 37% overlap, respectively (4,16)], whereas studies performed using the same platform were shown to be highly reproducible (12,16). Similarly, Guryev et al. (17) found a very good correspondence between WGS-based predictions and hybridization methods. On the other hand, changes in calling methods were shown to modify the number of CNV predictions, with methods that congruently analyze aCGH data from multiple samples significantly increasing the number of detectable CNVs (14,16).

General characteristics of the CNVs identified in human and model organisms are summarized in Table 1. Interestingly, the CNVs reported in D. melanogaster are one order of magnitude shorter (maximum 33 kb or 0.017% of the fly genome) than those found in mammals (up to 3 Mb or 0.12% of the mouse genome). A certain synteny has been observed among primates with 22% and 25% of chimpanzee and macaque
CNVs overlapping human CNVs (7,19). In addition, CNVs identified in multiple macaques were frequently observed in multiple human samples, suggesting the existence of hotspots for CNV (19). Indeed, structural genomic rearrangements are facilitated by repetitive sequences that provide the substrate for non-allelic homologous recombination (25). Consistently, CNVs are associated with segmental duplications (SDs) in mammals (2,4,7,8,11,12,19) and to centromeres and telomeres (17,26), structures also associated with SDs. For example, as much as 60% of the basepairs mapping inside C57BL/6J SDs were recently shown to be polymorphic among classical mouse inbred strains (14). In other words, this type of variation appears to be far more clustered in discrete regions in this model than in humans, probably reflecting the unnatural man-driven establishment of inbred lines. Of note, CNVs and single nucleotide polymorphisms (SNPs) cluster in different regions of the domesticated mouse genome (10,33). Consistently, Cutler et al. (10) found that CNV segments were significantly enriched among sequences with low and moderate SNP content. The artificially repeated brother–sister mating of mice over generations has eliminated recessive lethal alleles, but may have increased the fixation of potentially (slightly) deleterious CNV alleles. Consistently, significantly more CNVs were identified in classical inbred strains than in wild-caught Mus musculus domesticus mice (16).

CNV regions are not only polymorphic within a species, but spontaneous de novo rearrangements can occur even in an inbred population, as shown by recent studies of the classical C57BL/6J strain (13,16,34). Egan et al. and Henrichsen et al. identified and validated 38 and 39 CNVs in this inbred strain, respectively, 18 of which occurred through recurrent mutations (13,16). In addition, a single-copy and a duplication allele of the CNV spanning the Ide locus were found to be in Hardy–Weinberg equilibrium 14 years after their appearance in the Jackson Laboratory colony (34). These findings illustrate the highly dynamic nature of CNV regions and challenge the notion of the isogenicity of inbred strains (35).

BACs, bacterial artificial chromosome; PCR, polymerase chain reaction.

## Table 1. Genome-wide copy number variation (CNV) surveys in human and model organisms

<table>
<thead>
<tr>
<th>Model organism</th>
<th>Platform</th>
<th>Probe density</th>
<th>Resolution</th>
<th>% of genome covered by CNVs</th>
<th>Size range of CNVs</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>BAC arrays</td>
<td>1/17 kb</td>
<td>1 Mb</td>
<td>~1.49%</td>
<td>Up to 2 Mb</td>
<td>(2,8)</td>
</tr>
<tr>
<td>Human</td>
<td>Representational Oligonucleotide Microarray Analysis (ROMA)</td>
<td>1/17 kb</td>
<td>35 kb</td>
<td>~1.49%</td>
<td>700 bp to 3.19 Mb</td>
<td>(5)</td>
</tr>
<tr>
<td>Human</td>
<td>BAC arrays and SNP genotyping arrays</td>
<td>1/4.1 kb</td>
<td>1/4.1 kb</td>
<td>12%</td>
<td>Up to 6 Mb</td>
<td>(4)</td>
</tr>
<tr>
<td>Human</td>
<td>Bioinformatics and BAC arrays</td>
<td>1/36 bp</td>
<td>&lt;1 kb</td>
<td>~2%</td>
<td>Up to 35 kb</td>
<td>(20)</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Spotted PCR array</td>
<td>1/6kb</td>
<td>50 kb</td>
<td>Up to 10%</td>
<td>40–3000 kb</td>
<td>(13)</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Euchromatic genome tiling array</td>
<td>1/6kb</td>
<td>1 Mb</td>
<td>60% of segmental duplications</td>
<td></td>
<td>(14)</td>
</tr>
<tr>
<td>Mouse</td>
<td>BAC arrays</td>
<td>1/6kb</td>
<td>40 kb</td>
<td></td>
<td>32–710 kb</td>
<td>(19)</td>
</tr>
<tr>
<td>Mouse</td>
<td>Oligonucleotide arrays</td>
<td>1/5 kb</td>
<td>1.4%</td>
<td></td>
<td></td>
<td>(17)</td>
</tr>
<tr>
<td>Mouse (single strain C57BL/6J)</td>
<td>Representational oligonucleotide microarray analysis (ROMA)</td>
<td>1/17 kb</td>
<td>1 Mb</td>
<td></td>
<td>1/1 Mb</td>
<td>(7)</td>
</tr>
<tr>
<td>Mouse</td>
<td>Segmental duplication-specific oligonucleotide array</td>
<td>1/6kb</td>
<td>1 Mb</td>
<td></td>
<td>1/1 Mb</td>
<td>(7)</td>
</tr>
<tr>
<td>Rat</td>
<td>Bioinformatics and oligonucleotide arrays</td>
<td>1/5 kb</td>
<td>1.4%</td>
<td></td>
<td></td>
<td>(17)</td>
</tr>
<tr>
<td>Macaque</td>
<td>Oligonucleotide arrays</td>
<td>1/6 kb</td>
<td>1.4%</td>
<td></td>
<td></td>
<td>(17)</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>Human BAC arrays (2)</td>
<td>1/1 Mb</td>
<td>1 Mb</td>
<td></td>
<td></td>
<td>(7)</td>
</tr>
</tbody>
</table>
PHENOTYPIC CONSEQUENCES OF CNVS

With their prevalence, e.g. 10% of the mouse autosomal genome and 60% of its duplicated regions (14,16), CNVs should constitute significant contributors to intraspecific genetic variation. Consistently, an initial case of adaptive CNV alleles under positive selection was recently uncovered (36). Similarly, copy numbers of individual CNVs have been associated with diseases or susceptibility to diseases, either through dosage of a single gene (37–44), a contiguous set of genes [e.g. Williams-Beuren syndrome and infantile spasms (45,46), DiGeorge Syndrome (47), Smith-Magenis syndrome (48), Potocki-Lupski syndrome (49)] or allele combinations in the case of complex diseases, especially of the central nervous system. Recent studies in schizophrenia (50,51), autism (52,53) and mental retardation (54–56) revealed multiple disease-causing rearrangements, but also lead to the description of novel, clinically recognizable micro-deletion syndromes and recurrent rearrangements causing variable phenotypes (57–65). Structural rearrangements in three regions, namely 1q21.1, 15q11.2 and 15q13.3 were shown to be associated with both mental retardation and schizophrenia. A more detailed investigation of cases might shed light on a possible link between these conditions. Interestingly, while deletions at all three loci were linked to schizophrenia and related psychoses (50,51), the maternal duplication of the entire 15q11-q13 region was shown to cause autism spectrum disorder (ASD) (66). Similarly, duplication of band 1q21.1 was identified in patients with ASD, whereas both deletions and duplications were associated with mental retardation. These examples suggest that some conditions might not be associated with a specific number of copies of a particular CNV, but rather that the simple presence of a structural change at a given position of the human genome may cause perturbation in particular pathways regardless of gene dosage.

CNV has also been shown to induce phenotypes in non-mammalian vertebrates and in invertebrates. For example, copy number amplification of regions containing the mannose-binding lectin and the pfmdrl and gch1 genes influence the zebrafish susceptibility to bacterial infection and malaria parasite Plasmodium falciparum drug resistance, respectively (67,68).

FROM GENOTYPE TO PHENOTYPE—BRIDGING THE GAP

Phenotypic effects of genetic differences, both at the single-nucleotide and large-scale level, are supposedly brought about by changes in expression levels, either directly affecting the genes concerned by the genetic change, or indirectly through position effects or downstream pathways and regulatory networks (69,70). Thus, transcriptome analyses and single-gene expression studies seem appropriate proxies to study the consequences of CNV.

A first attempt to estimate the contribution of CNVs to gene expression variation found that in human lymphoblastoid cell lines, changes in copy numbers capture about one-fifth of the detected genetic variation (71). This type of approach, however, does not allow to follow the influence of CNVs in multiple tissues nor developmental stages, as an individual will have been the source of a limited number of cell lines, typically a skin fibroblast and a lymphoblastoid cell line or, more rarely, different cell lines from umbilical cord. This difficulty can be overcome using animal models, as recent analyses suggest that CNVs are also important contributors to their phenotypic variation (Table 1). In some cases given CNVs were shown to cause the same phenotype in a model organism and in human patients, further exemplifying the validity of models studies.

For example, Atim et al. (40) showed that lower copy number of the Fcgr3 gene predisposed both rats and humans to immunologically mediated glomerulonephritis. In addition to natural CNV, structural modifications can be engineered into model organisms, e.g. mice, and thus allow the study of different alleles of a CNV, as well as position effects, in a genetically homogenous background (72,73).

Two recent reports attempted to assess the effects of CNVs on tissue transcriptomes of rats and mice (16,17). Both studies report an over-representation of differentially expressed genes among CNV-mapping transcripts. Globally, a weak yet significant positive correlation was found between relative expression level and gene dosage (see examples in Fig. 1). This correlation was driven by a restricted number of genes, as only a fraction of genes comprised between 5% and 18% in function of the tissue and the rodent species analyzed showed a strong correlation between number of copies and relative expression levels (16,17) (see example in Fig. 1A and possible mechanism in Fig. 2B). In about two-third of the genes, however, the number of copies had no effect on relative expression levels in any tissue (see example in Fig. 1C), suggesting either dosage compensation mechanisms or the incomplete inclusion of regulatory elements in the deletion/duplication event (see possible mechanisms in Fig. 2). In addition, genetic imprinting may also modulate the expression at a CNV locus, in a parental allele-dependent manner (74). The first hypothesis is supported by the observation that expression of some genes correlated with gene dosage in some tissues but not in others (see examples in Fig. 1D and mechanisms in Fig. 2B and C), whereas the second is supported by studies of the quantitative collinearity of Hoxd genes, which showed that their expression levels are correlated with their rank in the Hoxd cluster (75) (see mechanism in Fig. 2D). Alternatively, the multiple copies of a gene might be mapping to different chromatin environments and thus be regulated differentially (Fig. 2E); indeed, aCGH does not allow discriminating tandem from non-tandem duplications. Furthermore, for 2–15% of the genes depending on the tissue and the rodent assessed, the recorded relative expression levels were significantly inversely correlated with CNV (16,17) (see examples in Fig. 1B). The mechanism of this effect is still poorly understood, but may be explained by two models summarized in Fig. 2F and 2G. In the first model, a negative correlation between number of copies and relative expression is explained by immediate early genes (IEG, Fig. 2F). These genes are initially expressed at levels proportional to their number of copies (Fig. 2A and B). These IEG induce directly or indirectly the expression of a repressor which, by a negative feedback loop, reduces or even abolishes the expression of the CNV gene (Fig. 2F).
In the second hypothesis, the extra copies of a gene impair through steric hinderance their access to a specific transcription factory, where this particular locus should be transcribed (76) (Fig. 2G).

Genes mapping to CNVs show lower expression levels than transcripts that do not vary in numbers of copies, both in fly and mouse (16,20). These observations suggest that CNV genes might be more specific in their expression patterns than other genes. In support of this notion, CNV genes were detectable in a smaller number of tissues than non-CNV genes both in the mouse and *Drosophila* (16,20). In addition, Dopman and Hartl (20) observed that the CNV genes expressed in the fly midgut and male accessory glands were enriched for functions relevant to these tissues, defense response in both cases and post-mating behavior genes in the latter case.

Large-scale rearrangements do not only affect expression by altering gene dosage. The deletion associated with Williams-Beuren syndrome was shown to modify the expression of some of the normal-copy number neighboring genes in human lymphoblastoid and skin fibroblast cell lines (77). Consistently, Stranger *et al.* (71) observed that although CNVs capture 18% of the detected genetic variation in lymphoblastoid cell lines, more than half of the identified associations between copy numbers and expression levels involved genes mapping outside the CNV intervals. Likewise, an engineered duplication of a segment of mouse chromosome 11, that models the rearrangement present in Potocki-Lupski syndrome patients,
was shown to affect the expression of both the genes mapping
within the duplicated interval and its flanks (72). Thus, CNVs
also profoundly affect the expression of genes located in their
vicinity. This effect extends over half a megabase into their
genomic neighborhoods (16,17). But how may changes in
copy number of CNV regions alter the expression of genes in
their vicinity? Different CNV-induced mechanisms that
include the physical dissociation of the transcription unit from
its cis-acting regulators (78,79), modification of transcriptional
control through alteration of chromatin structure (80–83) and
modification of the positioning of chromatin within the
nucleus and/or within a chromosome territory of a genomic
region (84,85) might play a role, both individually or in com-
bination. Copy number changes might also influence gene
expression through perturbation of transcript structure
(70,86,87). Similar mechanisms might explain how some path-
ways are perturbed without correlation with a particular number
of copies of a CNV (see above). Detailed investigations of the
different mechanisms by which CNVs influence gene
expression are warranted to shed light on how CNVs alter the
architecture of chromosomal segments and thus influence the
expression of genes.

CONCLUSION

The identification of CNVs in both human and model organ-
isms and their association with diseases and phenotypes
have confirmed that these genome structural changes are
important contributors to variation. They exert their influence
by modifying the expression of genes mapping within and
close to the rearranged region.

Conflict of Interest statement. None declared.

FUNDING

Jérôme Lejeune Foundation, the Telethon Action Suisse
Foundation, the Swiss National Science Foundation and the European
Commission anEUploidy Integrated Project (grant 037627).
REFERENCES


