Pathogenic or not? Assessing the clinical relevance of copy number variants


The availability of commercially produced genomic microarrays has resulted in the widespread implementation of genomic microarrays, often as a first-tier diagnostic test for copy number variant (CNV) screening of patients who are suspected for chromosomal aberrations. Patients with intellectual disability (ID) and/or multiple congenital anomalies (MCA) were traditionally the main focus for this microarray-based CNV screening, but the application of microarrays to other (neurodevelopmental) disorders and tumor diagnostics has also been explored and implemented. The diagnostic workflow for patients with ID is now well established, relying on the identification of rare CNVs and determining their inheritance patterns. However, experience gained through screening large numbers of samples has revealed many subtleties and complexities of CNV interpretation. This has resulted in a better understanding of the contribution of CNVs to genomic disorders not only via de novo occurrence, but also via X-linked and recessive inheritance models as well as through models taking into account mosaicism, imprinting, and digenic inheritance. In this review, we discuss CNV interpretation within the context of these different genetic disease models and common pitfalls that can occur when searching for supportive evidence that a CNV is clinically relevant.

Conflict of interest

The authors have no conflict of interest to declare.

It is well recognized that copy number variation (CNV) is not only a prevalent source of genomic variation in the general population, but also is a major cause of many developmental disorders including intellectual disability (ID) and congenital anomalies, and contributes to later onset common diseases such as schizophrenia and autism (1). The robust detection of CNVs using genomic microarrays, has resulted in the widespread application of microarrays as a first-tier diagnostic tool (2–4). The current generation of genomic microarrays consists of hundreds-of-thousands to millions of probes and can result in the detection of tens to hundreds of CNVs in a single individual, ranging in size from 1 to 10 kb to several megabases (5). Recent work from our group and that of others has shown that clinically relevant, dominant de novo CNVs can be identified in 10–20% of individuals with ID (3, 6–8). However, interpreting the pathogenicity of CNVs can remain challenging and relies heavily on frequency information from healthy control cohorts and databases with previously reported clinically relevant CNVs (9, 10). In addition to dominant, de novo CNV interpretation, we will discuss disease models including X-linked inheritance, recessive inheritance, mosaicism, imprinting, and digenic disease models. Many of these models can be more easily studied now that CNV detection has become a standard approach in genome diagnostics and is applied to large patient cohorts. This will further improve once CNV detection is combined with the detection of other types of genomic variation using whole exome and genome sequencing. Furthermore, we will highlight several common pitfalls when performing CNV interpretation.
Table 1. Public data resources for CNV interpretation

<table>
<thead>
<tr>
<th>Name</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common variants</td>
<td></td>
</tr>
<tr>
<td>DGV</td>
<td><a href="http://projects.tcag.ca/variation">http://projects.tcag.ca/variation</a></td>
</tr>
<tr>
<td>Clinically relevant variants</td>
<td></td>
</tr>
<tr>
<td>ECARUCA</td>
<td><a href="http://www.ecaruca.net">http://www.ecaruca.net</a></td>
</tr>
<tr>
<td>CAGdb</td>
<td><a href="http://www.cagdb.org">http://www.cagdb.org</a></td>
</tr>
<tr>
<td>Combination of common and clinically relevant variants</td>
<td></td>
</tr>
<tr>
<td>DECIPHER</td>
<td><a href="http://decipher.sanger.ac.uk">http://decipher.sanger.ac.uk</a></td>
</tr>
<tr>
<td>ISCA CNV</td>
<td><a href="https://www.iscaconsortium.org">https://www.iscaconsortium.org</a></td>
</tr>
<tr>
<td>Known disease genes</td>
<td></td>
</tr>
<tr>
<td>OMIM</td>
<td><a href="http://www.omim.org">http://www.omim.org</a></td>
</tr>
<tr>
<td>HGMD</td>
<td><a href="http://www.hgmd.cf.ac.uk">http://www.hgmd.cf.ac.uk</a></td>
</tr>
<tr>
<td>Gene imprint</td>
<td><a href="http://www.geneimprint.com">http://www.geneimprint.com</a></td>
</tr>
</tbody>
</table>

Dominant de novo

The clinical implementation of microarrays for CNV screening started in the cytogenetics arena where the majority of patients studied were sporadic cases affected by ID and/or multiple congenital abnormalities (MCA). As a result, CNV interpretation has traditionally focused on dominant disease models and the identification of rare de novo CNVs of considerable genomic size (Fig. 1d). A typical diagnostic CNV interpretation workflow starts by using frequency information from healthy control cohorts to discriminate between common (probably benign) CNVs and rare (possibly pathogenic) CNVs. Information on CNV frequencies in unaffected controls can be found in public databases (Table 1), however many clinical testing facilities also use local or national databases. These local CNV collections sometimes have improved capture of less abundant, population-specific and common CNVs. CNVs can be classified as benign when similar CNVs have been found in multiple (~>1%) unaffected controls. For CNVs larger than 100 kb, depending on the size of the control cohort, the number of potentially pathogenic CNVs can be reduced by approximately 90% (11). However, the use of control databases is less effective for smaller CNVs, only a few kilobases in size, which can be detected by newer generation, high-resolution microarrays and NGS technologies. This is a direct result of fewer control samples being available and some CNV regions being platform-specific due to probe coverage and microarray design. The submission of the CNV sets derived from large collections of whole exome and whole genome sequencing data to public databases will gradually overcome this issue.

Once a CNV has been identified as rare, several public databases are available which contain information on rare, possibly pathogenic CNVs (Table 1) (12, 13). The utility of these databases is dependent on both the number of cases that are (publically) available as well as the level of annotation and quality of the data submitted. In daily practice, many of these databases share a number of shortcomings, such as poor annotation of genomic and phenotypic data (platform information, clinical information and information of inheritance) and a lack of data curation (experiment quality information and CNV calling parameters) (14). As a result, third party databases should be used with care and common sense. Even after a decade of microarray CNV studies, it is important to realize that many CNVs identified in ID patients today have never been observed (or at least described) before, showing the enormous amount of private genetic variation present in our population (15). Hence, follow-up analysis of parental samples is required to prove that the CNV has either arisen de novo in the affected patient or segregates with the clinical phenotype in the family, as the rate of de novo CNVs in the general population is considered to be low, especially for larger CNVs (16–18).

In addition to frequency and inheritance information, other factors can be used to aid CNV interpretation such as the (i) size, (ii) location (interstitial/centromeric/repeat regions), (iii) type (copy number state) and (iv) gene content of the CNV (3, 10, 19–21). Since the size of a CNV is associated with the number of genes affected, the likeliness of pathogenicity increases with the size of the CNV. Previous reports suggest an odds ratio of 13.71 for CNVs larger than 1 Mb in ID-associated phenotypes in comparison to controls (22). In addition, there is a significant increase in CNV burden in individuals with ID and MCA in comparison with individuals with ID alone. This suggests that the total CNV burden positively correlates with the severity of childhood disability (23, 24). Moreover, a deletion of one or more genes is more likely to have a pathogenic effect than the gain of gene copies, hence CNV losses are also more likely to have a pathogenic affect than CNV gains (Fig. 1) (6).

Dominant inherited

Whilst the de novo paradigm for both CNVs as well as single nucleotide variants (SNVs) (25) has been established in ID because of the severe reduction in reproductive fitness in this disorder, this effect is less strong for genetic disorders such as autism and schizophrenia. Rare inherited CNVs are much more likely to contribute to the phenotype and segregate with disease in these cohorts (26), in addition to de novo CNVs observed in sporadic cases (27, 28). In contrast to individuals with ID, de novo CNVs can be found in 2–7% of individuals affected by autism depending on the number of affected siblings in the family. One of the most frequently observed rare CNVs in autism cohorts is a maternally inherited CNV 15q11-q13 accounting for 1–3% of all cases (1, 28).

Imprinting

When a rare CNV is encountered in a patient, but inherited from a healthy parent, the CNV is most likely to be categorized as variant of unknown significance (VOUS) or likely benign. However, an apparently healthy carrier could be a non-penetrant carrier of a
CNV. Alternatively, if the CNV contains an imprinted gene the phenotypic impact will depend on whether the gene is maternally or paternally inherited.

Genomic imprinting is an epigenetic form of gene regulation (silencing) that leads to parent-specific differential expression of a subset of genes (29). At present, 87 imprinted genes are known in the human genome and reported in Gene Imprint (Table 1), including a cluster of imprinted genes located on the proximal long arm of chromosome 15. Genomic imbalances frequently occur in 15q11q13, because it harbors numerous low-copy repeats (LCRs; BP1 through BP6) that mediate duplications, deletions or chromosome rearrangements through non-allelic homologous recombination (30). The $\textit{SNRPN}$ gene located on 15q11.2 is maternally imprinted in all tissues whereas the nearby $\textit{UBE3A}$ gene is paternally imprinted in the brain. A recurrent 6 Mb deletion of 15q11.2q13.2 (BP1-BP3) is known to cause Prader–Willi syndrome when the paternal 15q allele is deleted and Angelman syndrome when the maternal allele is deleted (31). In contrast, a 6 Mb duplication will have clinical consequences when maternal in origin, but will have no or limited phenotypic effect when the CNV is of paternal origin (32). Similarly, dysregulation of the 11p15 region can result in two different fetal growth disorders, Silver–Russell (SRS) or Beckwith–Wiedemann (BWS) syndrome, with opposite growth phenotypes. A duplication of the paternal 11p15 allele causes BWS, but duplication of the maternal 11p15 allele may cause SRS, depending on its size and breakpoints (33). Hence, the phenotypic consequences of CNVs in regions that harbor imprinted genes depend on whether the CNV has been maternally or paternally transmitted (Fig. 1g) and determining the inheritance patterns of the CNV is crucial for appropriate genetic counseling.

X-linked

On the basis of our own experiences, 10% of all potentially causative CNVs in ID are X-linked CNVs, of which 80–90% are maternally inherited (11). X-linked CNVs are often difficult to interpret requiring different criteria than autosomal CNVs (Fig. 1f) (34, 35). In females, the clinical consequences of an X-linked CNV may vary widely amongst individuals due to skewed X inactivation. In addition, the pattern of X inactivation may vary between different cell types (36). To assess the clinical relevance of a CNV on the X chromosome, carrier testing only needs to be performed in the mother for deletions in male patients, in all other cases both parents should be tested. If the same CNV detected in the patient is also seen in the father, the likelihood of the CNV being benign is high. If the CNV is maternally inherited the clinical relevance remains unclear and additional testing such as X-inactivation profiling and carrier testing of the maternal grandparents and/or uncles is required. When
no male carrier(s) of the same CNV are found, the pathogenicity of such CNVs may stay unclear.

Recessive inheritance

Individually, the majority of recessive conditions remain rare and the contribution of CNVs to carrier status and disease is not yet fully investigated. Nonetheless, evidence suggests that pathogenic alleles carrying CNVs in recessive disease genes occur frequently (37). Hence, not only the origin and direction of inherited CNVs should be considered, but also the specific copy number state. For example, a heterozygous loss in both parents, can result in a clinically relevant homozygous loss in the child (Fig. 1i). This was previously described by Knijnenburg et al. (38) who found a homozygous deletion in 15q13.3, a normal variation locus, in a patient with hearing loss, inherited from non-consanguineous, hemizygous carrier parents, and similarly for homozygous deletions affecting NPHP1 (39). If the CNV results in a gain, three copies of this genomic region may have no phenotypic effect, the parents being healthy carriers, but four copies may have clinical consequences for their child (40). Therefore, assessing the resulting net genomic material loss or gain in the child is important and CNV detection algorithms should distinguish between different CNV loss and gain states.

Alternatively, when one parent is a carrier of a CNV loss that contains a recessive disease gene and the other parent carries a mutation in the same gene, compound heterozygosity can lead to affected offspring (Fig. 1j). This was recently shown by Paciorkowski et al. (41) who reported on two unrelated patients with severe microcephaly, agenesis of the corpus callosum, scalp rugae, and a fetal brain disruption-like phenotype with inherited deletions of 16p13.11 in combination with a pathogenic mutation (nonsense and frameshift, respectively) in NDE1 on the non-deleted chromosome 16 homologue. Similarly, Rodríguez-Pascau et al. (42) described two novel 18q11q12 microdeletions in two Niemann-Pick type C disease patients who each had a mutated NPC1 gene on the other 18q allele. Another mechanism was recently elucidated by Albers et al. (43), who showed that the compound inheritance of a rare null allele and one of two low-frequency SNPs in the regulatory regions of RBM8A, encoding the Y14 subunit of the Exon Junction Complex, causes TAR syndrome. Assessing the pathogenicity of a CNV should therefore not only be limited to dominant disease effects, but should also include recessive disease models in combination with other forms of genomic variants resulting in loss of function (37).

Two-hit CNV model

Unbiased genomewide CNV analysis by microarrays has recently also provided the first insight into complex inheritance models in which the combination of two or more inherited CNVs can explain a phenotype. In 2010, Girirajan et al. (44) studied developmental delay patients with recurrent 16p12.1 microdeletions and observed an enrichment of additional large CNVs (>500 kb) in comparison to matched controls. Of note, the 16p12.1 deletion patients with additional CNVs showed a more severe clinical phenotype than those without additional CNVs. In a follow-up of this work, the same authors expanded this analysis to a series of more than 2000 patients carrying a CNV associated with ID and congenital malformations (45). In this analysis, 10% of patients were observed to carry additional large CNVs that may contribute to the phenotype. These additional CNVs occurred more frequently in addition to CNVs affecting 1q21.1 (46), 16p11.2 (47) or 16p12.1 (44) that are mostly inherited and known to show considerable phenotypic variability. In contrast, de novo CNVs causing more homogeneous (syndromic) clinical phenotypes are rarely associated with additional large CNVs. These are first observations of digenic disease models, and we will likely see many examples of digenic or oligogenic inheritance once whole genome sequencing becomes the standard approach to study all types of genomic variation in a single experiment.

Mosaic

Genomic microarray analysis is also very suitable for detecting mosaic genomic imbalances present in as few as 10–20% of cells. The presence of SNP probes on a microarray greatly aids in mosaic CNV detection since the allelic imbalances present in mosaic samples can be visualized in so called B-allele frequency plots. This additional information improves the detection of mosaic CNVs and also allows the detection of copy neutral events such as (segmental) uniparental disomy (UPD). Since a constitutional whole chromosome UPD is the result of either a monosomy or trisomy rescue (48), the UPD itself might not always be directly disease causing, but the initial aneuploidy may still be present in (low) mosaic form and the actual cause of the clinical phenotype (49, 50).

Moreover, a CNV may sometimes seem to be inherited from an apparently healthy parent, but additional testing can reveal that the parent is in fact a mosaic carrier of the same CNV that was encountered in their affected child (Fig. 1h). Kousoulidou et al. (51) recently reported on a patient with developmental delay, severe speech delay and facial dysmorphism with an intragenic TCF4 loss of 265 kb that he inherited from his clinically unaffected, mosaic father. An individual’s level of mosaicism may also vary over time and between cell types (52–54). The percentage of cells with the pathogenic CNV will determine the phenotype of the carrier parent ranging from normal to (mildly) affected. However, very low-grade mosaicism will remain undetectable when only one cell type is examined, therefore testing of other tissues or cell types should be considered when the presence of low-level somatic-gonadal mosaicism is suspected (55).
Pathogenic or not? Assessing the clinical relevance of CNVs

Pitfalls

Non-genic de novo

Evaluation of the gene content affected by a CNV is included in many interpretation workflows (3, 10, 19–21). Current knowledge about the possible effects of a CNV in a non-coding region is still very poor, making clinical interpretation of such regions and subsequent patient counseling difficult (Fig. 1c). Consequently, intergenic or intronic CNVs are often discarded during clinical interpretation and reporting for practical reasons. However, a number of examples of pathogenic CNVs in non-coding regions have been reported, such as a duplication 5′ of GRIA3 on Xq25 (56), deletions 3′ of PAX6 on 11p13 (23), and a deletion of the IHH locus at 2q35 (57), reviewed by Spielman and Klopocki (58). As our knowledge of the human genome expands, the clinical interpretation of these non-coding CNVs will become more feasible and CNV data sets could be re-analysed to rescue these discarded, non-coding CNVs. However, at present when assessing a CNV, the flanking genomic regions (±1 Mb) should always be checked for relevant (disease) genes that could be subjected by position effects, and one should remain critical, not stopping interpretation as soon as one genic CNV is found.

Non-pathogenic, de novo CNVs

Many CNVs detected using high-resolution microarrays remain private, making it difficult to link their functional relationship with the phenotype. Current CNV interpretation workflows suggest that de novo CNVs are considered clinically relevant in ID patients (3, 10, 19–21). It is, however, well known that de novo germline mutations do occur in the general population (59), and for CNVs it is known that these occur de novo at very low frequency (60). Few examples exist of large, de novo CNVs with no apparent phenotypic consequences. Precise estimates of the de novo CNV mutation rate varies, but it has been estimated to be 1 in 50 individuals (17, 18). It is likely that with the increase in CNV detection resolution more examples will be discovered in the normal population of small de novo CNVs and the estimated mutation rate for CNVs will be refined. Current examples of de novo CNVs not contributing to a clinically relevant phenotype include a patient with ID and an eye disorder, in whom a de novo 86.5 kb deletion was identified. However, the patient was later characterized with a CHARGE-like phenotype and a de novo mutation was identified disrupting the CHD7 gene. Likewise, a 250 kb de novo deletion affecting MACROD2, was identified in a patient with Kabuki syndrome. Further mutation screening revealed a de novo mutation in MLL2. In both cases the clinical phenotype of the patient could be explained by de novo point mutations and not by the de novo CNVs present in these patients (61), highlighting the importance of detailed phenotypic information when performing clinical interpretation of both CNVs and SNVs.

Gain-of-function

It is not only important to consider the number of genes affected by a CNV, but also the function and the expression profile of the genes disrupted. The presence of dosage-sensitive genes or proven disease genes can aid in assessing the possible pathogenic effect of a CNV. When proven autosomal dominant disease genes are located in a deleted region, it is highly likely that this CNV will have a deleterious effect. However, public disease gene data sources such as OMIM and the HGMD (Table 1) should be checked to confirm that etiology of the disease is based on gene haploinsufficiency and not on a gain-of-function or a dominant negative mechanism. In the latter two cases, a deletion of such a gene may not have the same (if any) deleterious effect. A well-known example of a deletion of a dominant disease gene is the 15q11.2 BP1-BP2 deletion encompassing the NIPA1 gene associated with autosomal dominant spastic paraplegia type 6 (SPG6). The most supported disease mechanism of NIPA1 mutations is a dominant negative one, but a gain-of-function has also been hypothesized (62). Deletions are very common, and do not result in SPG6, but are a possible susceptibility factor for ID and autism (6). Likewise specific point mutations in SETBP1 exert their phenotypic effect causing Schinzel–Giedion syndrome through either a gain-of-function or a dominant negative effect and not via loss of function (63). Hence the individuals’ phenotype and the function of the gene should be considered when assessing the pathogenicity of a CNV affecting a known disease gene.

Genomic position of the event

A CNV gain may have clinical consequences due to the increased number of gene copies present or through disruption of the genes around the breakpoints. Region-specific fluorescence in situ hybridisation (FISH) analysis in routine diagnostics has shown that the majority of gains detected via microarray analysis are due to a duplication, possibly affecting the expression of the gene(s) at or near the duplication breakpoint(s) (64). As a result, the phenotypic consequences of a CNV gain depend on whether the copy number gain is due to a duplication occurring in tandem, an insertion elsewhere in the genome or due to the presence of a supernumerary marker chromosome. For example, CNV gains which have been mediated by non-allelic homologous rearrangements may be accompanied by an inversion depending on the direction of the flanking low copy repeats (65, 66). In case of a de novo copy number gain, whether on the X chromosome or an autosome, it is always important to consider region-specific FISH analysis in order to be able to more accurately determine the recurrence risk (67, 68).

Future directions

In addition to traditional CNV interpretation workflows, several attempts have been made to predict the impact
of CNVs based on computation methods. These tools include classifiers based on structural and functional genomic features used to annotate each CNV such as Long Interspersed Nuclear Element (LINE) density and mouse knock-out phenotypes (15). Alternatively, a gene by gene based approach can also be used to create maps of haplosufficient (HS) and haploinsufficient (HI) genes. The study of such genes based on their genomic, evolutionary, functional, and network properties shows that HI genes are typically longer and have more conserved coding sequences and promoters than HS genes. HI genes exhibit higher levels of expression during early development and greater tissue specificity. Moreover, within a probabilistic human functional interaction network HI genes have more interaction partners and greater network proximity to other known HI genes (69). More recently, the Ensembl Variant Effect Predictor has been updated to annotate not only SNVs but also indels and CNVs based on their overlap with known regulatory features, location in transcription factor binding sites and whether the variant has been previously reported (70, 71). Whilst the application of such tools is limited in the clinic, the they can provide useful information for prioritizing CNVs.

The diagnostic implementation of NGS technologies such as whole exome and especially whole genome sequencing creates new possibilities for the simultaneous testing of both SNVs, indels and CNVs. While identification of all genomic variation in a single experiment is on the horizon, effective and confident clinical interpretation of all of this information will remain a considerable challenge. The only way forward is to standardize genotypic and phenotypic information, share data between both clinical centers and population studies in well-curated databases, and for each geneticist to ensure that diagnostics never becomes too much of a routine, each patient’s genome is unique and can teach us many useful lessons.

Acknowledgements
Dr Veltman is supported by the European Research Council (DENOVO 281964).

References
Pathogenic or not? Assessing the clinical relevance of CNVs


72. Neill NJ, Ballif BC, Lamb AN et al. Recurrence, submicroscopic complexity, and potential clinical relevance of copy gains detected by array CGH that are shown to be unbalanced insertions by FISH. Genome Res 2011: 21: 535–544.

