Published in final edited form as:

Methods. 2009 September ; 49(1): 18–25. doi:10.1016/j.ymeth.2009.06.001.

Copy Number Variants (CNVs) in primate species using array-based comparative genomic hybridization

Ömer Gökçümen and Charles Lee

Brigham & Women's Hospital and Harvard Medical School

Abstract

A substantial amount of genomic variation is now known to exist in humans and other primate species. Single nucleotide polymorphisms (SNPs) are thought to represent the vast majority of genomic differences among individuals in a given primate species and comprise about 0.1% of the genomes of two humans. However, recent studies have now shown that structural variation may account for as much as 0.7% of the genomic differences in humans, of which copy number variants (CNVs) are the largest component. CNVs are segments of DNA that can range in size from hundreds of bases to millions of base pairs in length and have different number of copies between individuals. Recent technological advancements in array technologies led to genome-wide identification of CNVs and consequently revealed thousands of variable loci in humans, comprising as much as 12% of the human genome (Jafrate1). CNVs in humans have already been associated with susceptibility to certain complex diseases, dietary adaptation, and several neurological conditions. In addition, recent studies have shown that CNVs can be successfully implemented in population genetics research, providing important insights into human genetic variation. Nevertheless, the important role of CNVs in primate evolution and genetic diversity is still largely unknown. This chapter aims to outline the strengths and weaknesses of current comparative genomic hybridization array technologies that have been employed to detect CNV variation and the applications of these techniques to primate genetic research.

Keywords

rhesus macaques; chimpanzees; array comparative genomic hybridization; microarrays; genomic evolution; structural variation; primate evolution

Copy Number Variation

Copy number variants are DNA segments that are duplicated or deleted in the genomes of organisms and vary in copy number among individuals. The importance of copy number variants (CNVs) was unknown for a long time, primarily because it was assumed that CNVs represented only a very small fraction of genomic variation. However, it was recently shown that CNVs are, in fact, surprisingly widespread in the genomes of phenotypically normal humans (2,3). Many subsequent studies have followed and now more than 6,000 CNV regions

© 2009 Elsevier Inc. All rights reserved.

Correspondence: Dr. Omer Gokcumen Cytogenetics Research Laboratory Department of Pathology Brigham and Women's Hospital Harvard Medical School 221 Longwood Avenue, EBRC 404 Boston, MA, USA 02115 Telephone: 617-278-0064 Fax: 617-264-5176 ogokcumen@partners.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

have been published for the human genome (1,⁴⁻⁷), covering approximately 12% of the human reference genome. Other studies demonstrated that CNVs are also a major factor in genomic variation among primate species (8-¹²). In addition, CNVs may result in different expression levels of specific genes (13-¹⁵), leading to differences in susceptibility to certain common diseases (16-²¹), as well as, evolutionary adaptations (22-²⁴). Therefore, the comprehensive analysis of intra- and inter-specific CNVs is a new and exciting opportunity to understand better the evolutionary history and phenotypic variation among primates.

The definition of CNVs has been a dynamic one. One established description of CNVs has been that they are intra- or inter-chromosomal duplications or deletions of 1 kb or larger DNA segments, but not represented by high copy number repetitive sequences, such as long interspersed nucleotide elements (LINEs), or pericentromeric tandemly repeated DNA sequences (25,²⁶). However, recent high-resolution genome maps have revealed smaller, yet common, CNVs among healthy humans (e.g., ^{6,27}), as well as complex structures within those CNVs (28). Furthermore, hundreds of thousands of insertion and deletion (INDEL) polymorphisms, ranging from 1 bp to 9989 bp in length have also been identified (e.g., ²⁹). Hence, from such recent findings, the traditional definition of CNVs is often expanded to include gains and losses that are a few hundred bases and larger. For this review, we will focus on consider CNVs that can be studied with current aCGH, which are generally equal or larger than 500 bases in length.

Our understanding of the genomic and cellular mechanisms for CNV genesis and evolution has also diversified in recent years. Currently, non-allelic homologous recombination (NAHR), which involves crossing-over between highly homologous DNA segments during meiosis, creating deletions or duplications in the homologous chromosomes, is the best studied mechanism through which larger CNVs can be generated (30). Another well studied category of CNV genesis is the size variation within variable number tandem repeats (VNTRs), which are short tandemly organized DNA sequences that are repeated in variable number between individuals (31). In addition to those, non homologous end joining (NHEJ), which is a highly conserved double stranded DNA repair mechanism, is discussed as a mechanism for CNV generation during G1 or G2 of the cell cycle (32-³⁴). More recently, fork stalling template switching (FosTeS) during mitosis (33,^{35,36}), and simple allelic recombination during meiosis (37) have been discussed as another mechanisms for generation of CNVs. The details and implications of these mechanisms are described elsewhere (38-⁴¹).

Among several methodologies (42-⁴⁶) utilized to identify copy number variation, comparative genomic hybridization (CGH) is one commonly used method for genome-wide CNV detection and genotyping. CGH arrays work by co-hybridization of differentially-labeled reference and sample DNAs to carefully selected and well-defined probe sequences. However, most commercially available CGH arrays are designed with human probes, and they are problematic when studying nonhuman primates because of cross-species sequence mismatches between the human probes and the orthologous DNA sequences in nonhuman primates. Designing custom species-specific probes, based on published full genome sequences of nonhuman primates, can overcome this problem, assuming that the primate reference genomes obtained are well annotated. Currently, full-genome sequence data are available for only a handful of primates. Nevertheless, it is safe to assume, considering the advent of next-generation sequencing technology, that the whole genome sequences of many more new primate species will be available within a couple of years. Therefore, the focus of the primate CNV research will likely shift from identifying and genotyping primate copy number variation to interpreting, at the population level, the evolutionary and population events that shape the structural genomic variation within and among primate species.

In this article, we aim to outline the applications of current array technologies that are employed for CNV genotyping and detection within and among primates. In short, we discuss the potentials and pitfalls of designing CGH array experiments to further our knowledge on structural genomic variation, and review the methodological interpretative challenges that lay ahead.

Experimental Design

As mentioned earlier, array CGH involves competitive hybridization of reference and test DNA samples (differentially labeled with cy3 and cy5) to previously designed probes¹ that are “spotted” on glass slides (Figure 1). The resulting hybridization signal ratios are measured for each probe (generally as $\log_2$ ratios) and significant deviations from zero indicate variations in the copy number of that particular DNA segment in the test sample, relative to the reference. Despite the increasing resolution of current array CGH technology, the copy number calls made with these newer array platforms still prone to false CNV calls especially for smaller and GC-rich CNVs. In addition, while higher-resolution array CGH platforms can narrow down the breakpoints that define a given CNV, they usually cannot identify the breakpoints to the nucleotide level. Nevertheless, array CGH remains a cost-effective, comprehensive and high-throughput means for accurate genome-wide detection and genotyping of CNVs.

The first and arguably the most important step for array CGH studies is the probe-design. For studies involving primates, three main issues other than the standard chemical hybridization quality controls (i.e., T_m , GC content, etc.) should be considered: species specificity of the probes, the inclusion (or exclusion) of repeat content, and the coverage of the customized platform. Most commercially available array CGH platforms are designed with probes that are based on the human reference genome (Figure 2). Hence, almost the entirety of our knowledge about whole-genome variation in non-human primates has been obtained from studies that utilize human probes (e.g., ^{10,11,47}) (also see Figure 3). For instance, the same BAC-based probes initially used to determine the widespread copy number variation in humans (3) were later utilized to detect copy number variation among 20 unrelated chimpanzees, resulting in the identification of more than 250 CNVs in chimpanzees, as well as “hot spots” for CNV formation (10). A similar study, which utilized a higher-resolution human CGH array for detecting intra-species copy number variation of 30 humans and 30 chimpanzees, revealed 353 autosomal CNV regions in humans and 438 in chimpanzees (11).

Despite the success of these initial studies with chimpanzees, it should be noted that for studies involving nonhuman primates, it is crucial to design custom, primate-specific probes to eliminate hybridization bias due to sequence mismatches with human probes. The influence of such bias will be greater in species that are evolutionarily more distant to humans. Attempts in our laboratory to identify CNVs with human arrays among species that are closely related to humans, such as orangutans and gorillas, have produced noisy, and generally irreproducible results. High copy number gains or homozygous deletions were more confidently identified, but in general, arrays designed from human DNA sequences can be problematic for studying CNVs in nonhuman primates (other than chimpanzees) due to substantial sequence divergence (see Figure 4). In addition to single base substitution mismatches, species-specific gains or losses (or human lineage-specific deletions) will not be represented in the human-specific CGH arrays. In addition, the sequence divergence between species significantly influences the hybridization signals obtained (48). In other words, even if the non-human DNAs hybridize to human probes, the sequence mismatches may influence the signal intensities, and consequently the observed $\log_2$ ratios.

¹Even though BAC clones are still commonly used, most commercially available companies now use oligo-based probes.

The only commercially available catalogue array for non-human primates is the 385K Roche-Nimblegen platform for rhesus-macaques (8). Other than this, Roche-Nimblegen Systems and Agilent Technologies, the leading companies that design/print CGH arrays, only provide custom-design services, through which species-specific arrays can be designed for nonhuman primates. For instance, Lee et al. (2008) designed a macaque-specific oligo-based CGH array based on the Rhesus macaque genome assembly available at the time (8). This study successfully revealed 123 CNVs among 10 unrelated macaque individuals and confirmed the previous observation that there are certain regions of the genome (i.e., those associated with ancestral segmental duplications) that serve as hot spots for CNV formation. A similar approach can therefore be taken to identify and genotype CNVs in chimpanzees and marmosets, two other primate species for which reference genome assemblies are now available.

With the advent of next-generation sequencing technologies, it is likely that there will be a dramatic increase in the number of primate whole genome sequences that will be available. Until then, one viable option for studying the copy number variation in these species is to design probes based on the sequences that are common in all primates for which the reference genomes are available (humans, macaques, chimpanzees, and marmosets). The coverage of these probes may be somewhat limited to regions where there is a high level of sequence-level conservation, which may create coverage problems, especially given that the mammalian conserved elements are depleted in regions where copy number variation was observed in humans (49). In addition, lineage-specific gains and species/lineage-specific regions under high selective pressure, which may have very important evolutionary implications, may be completely missed. Nevertheless, such limitations, “cross-primate” arrays may be useful in evolutionary studies focusing on the copy number variation in conserved regions of the genome.

Repeat Content

The large repeat content of the genome poses another set of challenges in designing copy number variation experiments. Most commercially-available platforms are repeat-masked, i.e., they do not cover most of the repetitive regions in the genome. For instance, the Agilent platform uses a similarity filter in determining their first tier probe sets that they use for their catalogue arrays (and for most of their custom arrays). This filter eliminates any sequences that align more than once in any part of the genome. However, it has been shown in recent studies that repeat segments may play fundamental roles in CNV generation and evolution (Conrad et al. unpublished results). Given the highly polymorphic nature of the repeat elements and their widespread distribution across chromosomes (especially that of mobile elements), designing probes for repetitive regions presents significant problems, and consequently many studies have altogether ignored the repeat content of the genome in their analyses of copy number variation.

Segmental duplications (SDs) should be considered under a different light within the context of designing probes especially for interspecies CNVs. Several studies have shown that CNVs are significantly enriched for SDs, which are large (>1kb), highly homologous (>90% sequence identity) DNA segments (1,¹⁰,⁵⁰⁻⁵³). In fact, non-allelic homologous recombination (NAHR), which involves SDs and may result in gain or loss of genetic material, has been discussed as one of the main mechanisms for CNV genesis (34,³⁸,⁵⁴). Hence, SDs should be incorporated into the studies of copy number variation among primates as some SDs are actually copy number variable between species and, from an evolutionary perspective; they are essentially CNVs and when fixed in a species, revert to SDs. It has been shown that within the evolutionary time, SDs facilitate the genesis of CNVs, and also other SDs (34). Hence, it is not surprising that throughout the primate lineage these regions become hot spots for CNV genesis (10). In practice, it is advisable to incorporate probes that fall into SDs, which essentially requires

loosening of the standard similarity filters that eliminate sequences aligning to multiple chromosomal loci within the genome.

A considerable number of currently documented CNVs contains large blocks of tandem repeats and/or mobile elements (31). Mobile elements are a major component of genomes of primates, comprising 40-50% of their genomic content. They are useful tools for studying the phylogenetic relationship between and within primate species, even though it has been shown that their activity is slowed within hominid lineages (55). SINEs, and Alu elements in particular, are shown to have very weak associations with primate CNVs (34). Long interspersed elements, on the other hand, are enriched in and around CNVs, supposedly since they are capable of facilitating NAHR (30). Based on these properties, some researchers try to incorporate LINEs in their studies of copy number variation in primates. However, then the challenge becomes to locate probes that can discriminate between dozens of highly homologous repeats dispersed across chromosomes and to distinguish the mobile elements of interest at each chromosomal location. One option is to design probes using sequence differences between LINE elements to increase the specificity of the probes. This approach, however, requires detailed maps of sequence diversity among LINEs, which is preliminary at this stage of the genome assembly. Overall, studying mobile elements that are associated with CNVs in the whole-genome level is at best a daunting task.

VNTRs, which include microsatellites and simple repeats, are particularly important to consider in studies of primate structural variation since they differ significantly across species. They have been shown to be divergent among primate species both in repeat number and in organization, and several studies have already utilized variation of the number of tandem repeats for understanding the genomic phylogenies (56,⁵⁷). For instance, the distribution and organization of functionally important alpha-satellites, which are sequence motifs that are repeated millions of times at centromeres, have been shown to vary significantly among primate species (58), especially between great apes and old world monkey species (59). However, most of the tandem repeat regions are poorly characterized at the sequence level in primate reference genomes. Therefore, designing DNA microarrays for these regions is difficult as most commercially-available array platforms mask out most of the tandem repeat regions in the genome.

Array slide experimental design and processing

Designing arrays for CNV studies greatly depends on the study question. For example, studies that aim to genotype many CNVs at once are usually targeted towards known CNV regions, especially if these regions are deemed functionally significant. Studies that aim to identify novel CNV regions often utilize probes that are selected at regular intervals throughout the genomes. For these arrays, the increased number of probes used to interrogate a given specimen correspond to a general increase in effective resolution in identifying and characterizing CNVs. The quality of the probes (related to the stringency of the probe designs) and the printing technology, among other factors, also influence the “effective” resolution of the arrays. In addition, Agilent employs post-scanning “extraction” of the array data, which eliminates significant amount of noise and reduces the deviation of signal intensities. For instance, our experience has shown that the effective resolution of 1 million probe Agilent arrays are roughly similar to that of 2.1 million probe Nimblegen arrays.

The experimental work associated with processing oligo-based CGH arrays has been refined and is very straightforward, albeit slightly different across different array companies. The first step usually involves the labeling of a reference and sample DNA with fluorescent dyes having different excitation and emission wavelengths (Cy3 and Cy5) and hybridization of the differentially-labeled DNAs to the array slide (Figure 1). As the probe density of the arrays

increases to millions, the required DNA for each sample (and reference) increases to as much as 2µg. One common way to reduce the false-positive errors is by performing experiments in a dye-swap strategy. In this experimental method, the reference and sample are processed twice, with different combination of dyes and the results of these two experiments should be complimentary.

Sample selection and choosing the appropriate reference sample constitute a crucial part of the experimental design. Since, all the CNV data obtained will be relative to this reference, it is essential to use high-quality, well-characterized reference DNAs. There are commercially-available pooled human genomes that some groups use as their reference sample. However, it is advisable to use a single, well-characterized genome as the reference for identifying intra-specific copy number variation. For instance, a recent study utilized single human and chimpanzee genomes for identifying CNVs within humans and chimpanzees, respectively (11). For studies that aim to determine inter-species copy number differences (CNDs), it may be more adventurous to use a reference DNA source containing many individuals (pooled DNA). Such a reference DNA source will provide an “average” copy number for each region, against which the species-specific variants can be delineated. It goes without saying that having a good knowledge of the absolute copy number states of the chosen reference(s) considerably increases the ability to interpret the aberrations in the reading of the arrays.

After scanning the hybridized slides, feature extraction and normalization is performed. Features extraction is the process of extracting of the intensities of each probe. During and after extraction, normalization is performed to reduce the systematic bias on the data, such as experimental errors or hardware malfunctioning. Both feature extraction and normalization requires sophisticated spot-recognition and image-processing software (e.g., Agilent Feature Extraction or Nimblegen NimbleScan). Post-scan image processing significantly reduces the background noise, dye-bias, and other anomalies stemming from non-biological variables and is an important part of array CGH data processing. The measured intensities, which are associated with specific chromosomal locations, can then be used to analyze the distribution of the CNVs among study samples. Array providers have their specialized software to process the scanned array images, incorporating feature extraction and normalization processes. There are also other algorithms and methods used for the second step, normalization, such as *qspline*. When using human probes with nonhuman primates, it is important to include additional steps to correct for the bias created by hybridization efficiency due to sequence mismatches.

Analyzing the data

CNVs are detected from the distribution of log₂ intensity ratios obtained from the reference and test hybridizations. The gains in the test sample will give positive log₂ ratios, whereas the losses will give negative values. Several software packages, including those from the array providers, utilize different algorithms to estimate the consecutive deviations of the log₂ intensity ratios to identify CNVs within the sample. The different algorithms involved take into account either some or all of the following parameters: the variation within the regions containing significant deviations of log₂ ratios, the relative log₂ ratios of the probes adjacent to each other, the signal-to-noise ratio of the overall array, the number of consecutive measurements that significantly deviate from zero, the posterior probability of CNV occurrence, among others. To date, the effective resolutions of the arrays allow detection of hundreds of CNVs in each individual. Each CNV call that is made computationally can also be checked manually. This is especially important when detecting CNVs that have complex structures, i.e., have smaller CNVs within larger CNVs (Figure 5), multi-allelic CNVs and copy number variants that have different breakpoints. However, the foreseeable exponential increase in the resolution of the arrays and the emergence of CNV-targeted arrays will push

the observed numbers of CNVs in each sample from hundreds to thousands, making manual checks less feasible. Therefore, more reliable algorithms and quantitative methodologies will be soon required for high-resolution, high-throughput studies of copy number variation both in humans and nonhuman primates.

The distribution of average log₂ ratios underlying an identified CNV region can be used to predict the copy number states of each individual, especially in intra-species experiments. This approach to genotype CNVs works best for biallelic CNVs, which would give three discrete clusters of log₂ intensity ratios obtained from population-based array CGH studies. However, determining the copy number states for multiallelic CNVs in each individual sample is challenging and error-prone, especially if solely based on distribution log₂ intensity ratios (Figure 6). The cross-species analyses of copy number difference can be done by comparing the two intra-species CNVs within the same orthologous region to delineate the ancestral state of the CNV. This latter comparison requires successful detection of the copy number states at the orthologous genomic locations of the compared species.

Validation

Validation is a key step in all array-based copy number variation studies to minimize the false-positive copy number calls. This issue gains additional prominence in studies involving nonhuman primates, given the lower accuracy of their reference genome assemblies (hence the reduction in the probe quality) leading to unavoidable probe mismatches with nonhuman primate DNAs on human-based arrays. There are several complementary methods to validate the CNVs identified by the array technologies, including polymerase chain reaction (PCR), quantitative PCR (qPCR), digital PCR, and fluorescence in situ hybridization (FISH). Moreover, there have recently been a flurry of new technologies, such as next-generation sequencing and mass spectrophotometry-based methods, which also show considerable promise for estimating the copy number states of samples from genomic DNAs accurately.

PCR using primers that are specifically designed to align flanking regions of smaller CNVs may be used to validate and/or estimate absolute copy number of the targeted loci. This method would be applicable in particular to deletions and tandem-duplications smaller than 5 kb. The size and, in cases of heterozygosity (i.e., different CNVs in homologous chromosomes) number of the observed bands may indicate the copy number state of the amplified DNA segments (Figure 7). Furthermore, the amplified DNA segments can then be sequenced directly to get nucleotide-level information on the CNV breakpoints of interest. Having the breakpoint information can provide considerable insights into the genesis and evolution of the CNVs, in particular by linkage disequilibrium (LD) analysis between SNPs and CNVs. Other more sophisticated PCR based approaches have also been employed for genotyping larger CNVs (e.g., ⁶⁰). However, PCR may be a limited tool for genotyping large, non-tandem and/or multiallelic CNVs. A more commonly used method for CNV validation is quantitative PCR (qPCR). Specifically, qPCR measures the amount of DNA amplification that occurs over a given time, which reflects the initial DNA quantity available at the start of the PCR reaction, with respect to reference genomic locus.

FISH is another important technique for identifying and characterizing individual CNVs (46). In this technique, fluorescent probes are designed to hybridize to specific chromosomal loci (in this case to CNV regions) and are then used to highlight regions of interest on chromosomes (or interphase nuclei). The chromosomes (or relaxed chromosomal fibers in *fiber-FISH*) are fixed to a glass slide and the probes are applied to the slides for hybridization. Metaphase or interphase FISH is especially important to localize CNVs that are not tandemly arranged, but are located at different chromosomal locations, or even on different chromosomes (23). It should be noted that one limitation of FISH is the requirement of viable cells for metaphase

chromosome FISH experiments. When frozen or paraffin embedded tissue samples are being used, one is limited to interphase FISH or fiber FISH.

Some emerging technologies also offer new solutions for validation. For instance, next-generation sequencing creates massive amounts of short sequence reads. From this data, gains and losses can be identified by several bioinformatic strategies. For example, read depth of contigs from this data can provide estimates of copy number states of each region. In addition, paired-end mapping algorithms and identification of split reads could be used in identifying CNVs in next-generation sequencing data. Even though next-generation sequencing technologies are expensive and methodologically cumbersome for routine copy number validation at this point, obtaining data of CNVs at the nucleotide level is a major advantage of this technology.

Interpretation

Recent association studies have attempted to associate copy number variation with expression profiles, diseases, and known phenotypic differences. However, as well-validated CNV data continues to accumulate for different species, the population-level and cross-species analyses of CNVs and CNDs will become an important venue for understanding selective pressures as well as demographic histories. Delineating the underlying genomic mechanisms that shape the current distribution of CNVs in primates will also be a topic of interest. It will also be important to describe the linkage disequilibrium (LD) of CNVs with other genomic variants, such as SNPs.

Most quantitative tools that are used in genetic association studies have been designed for SNPs. These tools incorporate comparative statistics, varying from basic t-test and regression to specifically designed sophisticated statistical algorithms, to assess whether a particular variant is prevalent among individuals who exhibit a particular trait, disease, or phenotype (61-63). Using this approach, CNVs have already been identified that appear to be associated with susceptibility and progression of several diseases, including HIV, Crohn's disease, and autism (20, 39, 64, 65). In addition, a large study of expression profiles among HapMap and other populations revealed associations of particular expression rates and certain CNV variants (13, 66, 67). Similarly, different copy numbers of salivary amylase gene (*AMY1*) have been shown to lead to differential expression of this gene, and the populations with higher starch consumptions tend to have higher copy numbers of *AMY1* (23). This study clearly demonstrated the adaptive advantage of a particular CNV within human populations with high starch consumption. Moreover, inter-species CNDs can be associated with phenotypic variation across species and can be important in understanding evolutionary divergence and functional differences between primate species.

The population genetics of copy number variation is still in its infancy, since the characteristics of most known CNVs are still ill-defined. In addition, there is still no reliable outgroup data for many comparative analyses (68). Despite these drawbacks, human copy number variation has already been shown to exhibit similar structural patterns with that of SNP variation across continents (69). Hence, similar to SNPs, most of the CNVs have arguably evolved under neutral conditions. In addition, preliminary measures of population differentiation involving log2 ratios and frequencies associated with CNVs, such as V_{st} , have been developed for data produced from array-based platforms (1, 69). Population genetic insights into copy number variation are crucial in delineating the demographic histories, selective pressures, and genomic mechanisms (i.e., mutations) that lead to copy number variation.

One important step in understanding the evolution of copy number variation among primates is to delineate the ancestral states of the CNVs. That is daunting, as the mechanisms that lead

to CNV genesis and evolution remain poorly understood. Under the assumption that there is no selection acting on CNVs, the CNV state that is shared by one or more species could be regarded as the ancestral state. For instance, if a CNV region varies between two or three copies among human populations and the orthologous region in chimpanzees invariably has two copies in all individuals, the neutral evolutionary model would suggest that the ancestral copy number state for this region is two. Such analyses can be extended to incorporate multiple species and used to build phylogenies. However, several CNVs and CNDs have already been shown to be related to phenotypic variation and evolved under selective pressures (9,¹¹).

In humans, it is estimated that ~23% of currently known CNVs reside in known or putative genes. Many CNVs that are outside genic regions of the genome may still have a significant influence on gene expression by effecting gene regulatory elements (66). Another important area of exploration is LD, i.e., non-random association between alleles at different loci, between CNVs and other genomic variants (especially SNPs), which may give important insights on the genomic evolution (e.g., the recombination patterns) of CNVs (61,⁶²). However, studies of LD between CNVs and SNPs is hindered by the fact that CNVs generally reside in genomic regions associated with segmental duplications and/or repeat-rich regions, which are difficult to sequence and detect SNPs. Moreover, the difference between CNV and SNP evolutionary rates are yet to be determined.

To complicate the studies of CNV evolution even further, recent studies have shown that there are hot-spots for CNV genesis (8,¹¹), and, in fact, that CNVs that are observed in multiple primate species are recurrent events within these structurally unstable regions. To delineate the distribution and structure of these regions will be important to delineate the mechanisms leading to CNV genesis, but also to better understand the CNV distribution within and among species. Indeed, more intra- and inter-species data are needed to delineate CNVs' relationship with other genomic variants, and to understand their distribution and categorization among primates.

Conclusion

This review summarizes some of the critical steps in designing, conducting and interpreting copy number variation studies in nonhuman primates using CGH arrays. Oligo-based CGH arrays allow very high-resolution genome wide structural variation surveys in humans, and the recent availability of customizable arrays have made possible to design species-specific arrays for nonhuman primates, for which the reference genome assemblies exist (humans, macaques, and marmosets). Human arrays have already been utilized to study chimpanzee CNVs. However, for studying the copy number variation of other primates, human-based arrays is inadequate due to the high sequence divergence. CNVs are an important part of the genetic variation in primates and can be associated with phenotypic variation. In addition, similar to SNPs, CNVs can give insights into demographic and evolutionary history of organisms. Ultimately, the accumulation of intra-species data will increase the interest in delineating the copy number differences between species. Such cross-species studies may shed light on the evolutionary history and mechanisms that shape the current primate genetic and phenotypic diversity.

References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Nature 2006;444:444–54. [PubMed: 17122850]

2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. *Science* 2004;305:525–8. [PubMed: 15273396]
3. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. *Nat Genet* 2004;36:949–51. [PubMed: 15286789]
4. Kang TW, Jeon YJ, Jang E, Kim HJ, Kim JH, Park JL, Lee S, Kim YS, Kim JY, Kim SY. *BMC Genomics* 2008;9:492. [PubMed: 18928558]
5. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM. *Nat Genet* 2006;38:86–92. [PubMed: 16468122]
6. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatiyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. *Nature* 2008;453:56–64. [PubMed: 18451855]
7. Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L, Aburatani H, Jones K, Redon R, Hurler M, Armengol L, Estivill X, Mural RJ, Lee C, Scherer SW, Feuk L. *Nat Genet* 2006;38:1413–8. [PubMed: 17115057]
8. Lee AS, Gutierrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, Miller GM, Korbel JO, Lee C. *Hum Mol Genet* 2008;17:1127–36. [PubMed: 18180252]
9. Degenhardt JD, de Candia P, Chabot A, Schwartz S, Henderson L, Ling B, Hunter M, Jiang Z, Palermo RE, Katze M, Eichler EE, Ventura M, Rogers J, Marx P, Gilad Y, Bustamante CD. *PLoS Genet* 2009;5:e1000346. [PubMed: 19165326]
10. Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Caceres AM, Iafrate AJ, Tyler-Smith C, Scherer SW, Eichler EE, Stone AC, Lee C. *Proc Natl Acad Sci U S A* 2006;103:8006–11. [PubMed: 16702545]
11. Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurler ME, Tyler-Smith C, Eichler EE, Carter NP, Lee C, Redon R. *Genome Research* 2008;18:1698–710. [PubMed: 18775914]
12. Kehrer-Sawatzki H, Cooper DN. *Hum Genet* 2007;120:759–78. [PubMed: 17066299]
13. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurler ME, Dermitzakis ET. *Science* 2007;315:848–53. [PubMed: 17289997]
14. Borel C, Antonarakis SE. *Mamm Genome* 2008;19:503–9. [PubMed: 18787897]
15. Mileyko Y, Joh RI, Weitz JS. *Proc Natl Acad Sci U S A* 2008;105:16659–64. [PubMed: 18946033]
16. Abrahams BS, Geschwind DH. *Nat Rev Genet* 2008;9:341–55. [PubMed: 18414403]
17. Bae JS, Cheong HS, Kim JO, Lee SO, Kim EM, Lee HW, Kim S, Kim JW, Cui T, Inoue I, Shin HD. *Biochem Biophys Res Commun* 2008;373:593–6. [PubMed: 18601904]
18. Colobran R, Casamitjana N, Roman A, Faner R, Pedrosa E, Arostegui JJ, Pujol-Borrell R, Juan M, Palou E. *Genes Immun*. 2009
19. Conrad B, Antonarakis SE. *Annu Rev Genomics Hum Genet* 2007;8:17–35. [PubMed: 17386002]
20. Schaschl H, Aitman TJ, Vyse TJ. *Clin Exp Immunol*. 2009
21. Milanese M, Segat L, Arraes LC, Garzino-Demo A, Crovella S. *J Acquir Immune Defic Syndr*. 2009
22. Patin E, Quintana-Murci L. *Trends Ecol Evol* 2008;23:56–9. [PubMed: 18191277]
23. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC. *Nat Genet* 2007;39:1256–60. [PubMed: 17828263]
24. Nei M, Niimura Y, Nozawa M. *Nat Rev Genet* 2008;9:951–63. [PubMed: 19002141]
25. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurler ME, Carter NP, Scherer SW, Lee C. *Genome Research* 2006;16:949–61. % U <http://www.ncbi.nlm.nih.gov/pubmed/16809666>. [PubMed: 16809666]
26. Feuk L, Carson AR, Scherer SW. *Nat Rev Genet* 2006;7:85–97. [PubMed: 16418744]
27. Korbel JO, Urban AE, Grubert F, Du J, Royce TE, Starr P, Zhong G, Emanuel BS, Weissman SM, Snyder M, Gerstein MB. *Proc Natl Acad Sci U S A* 2007;104:10110–5. [PubMed: 17551006]

28. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C. *Am J Hum Genet* 2008;82:685–95. [PubMed: 18304495]
29. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. *Genome Res* 2006;16:1182–90. [PubMed: 16902084]
30. Bailey JA, Eichler EE. *Nat Rev Genet* 2006;7:552–64. [PubMed: 16770338]
31. Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G. *BMC Genomics* 2008;9:533. [PubMed: 18992157]
32. Arlt MF, Mulle JG, Schaibley VM, Ragland RL, Durkin SG, Warren ST, Glover TW. *Am J Hum Genet.* 2009
33. Hazkani-Covo E, Covo S. *PLoS Genetics* 2008;4:e1000237. [PubMed: 18949041]
34. Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB. *Genome Res* 2008;18:1865–74. [PubMed: 18842824]
35. Lee JA, Carvalho CM, Lupski JR. *Cell* 2007;131:1235–47. [PubMed: 18160035]
36. Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. *Nature* 2005;437:94–100. [PubMed: 16136133]
37. Abu Bakar S, Hollox EJ, Armour JA. *Proc Natl Acad Sci U S A* 2009;106:853–8. [PubMed: 19131514]
38. Gu W, Zhang F, Lupski JR. *Pathogenetics* 2008;1:4. [PubMed: 19014668]
39. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. *Genomics* 2009;93:22–6. [PubMed: 18822366]
40. Korbel JO, Kim PM, Chen X, Urban AE, Weissman S, Snyder M, Gerstein MB. *Curr Opin Struct Biol* 2008;18:366–74. [PubMed: 18511261]
41. Feuk L, Marshall CR, Wintle RF, Scherer SW. *Hum Mol Genet* 2006;15(Spec No 1):R57–66. [PubMed: 16651370]
42. Plagnol V. *Hum Genomics* 2009;3:191–4. [PubMed: 19164094]
43. Beaudet AL, Belmont JW. *Annu Rev Med* 2008;59:113–29. [PubMed: 17961075]
44. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L. *Nat Genet* 2007;39:S7–15. [PubMed: 17597783]
45. Emanuel BS, Saitta SC. *Nat Rev Genet* 2007;8:869–83. [PubMed: 17943194]
46. Carter NP. *Nat Genet* 2007;39:S16–21. [PubMed: 17597776]
47. Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE. *Genome Res* 2003;13:347–57. [PubMed: 12618365]
48. Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP. *Genome Res* 2005;15:674–80. [PubMed: 15867429]
49. Derti A, Roth FP, Church GM, Wu CT. *Nat Genet* 2006;38:1216–20. [PubMed: 16998490]
50. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, Alkan C, Aksay G, Girirajan S, Siswara P, Chen L, Cardone MF, Navarro A, Mardis ER, Wilson RK, Eichler EE. *Nature* 2009;457:877–81. [PubMed: 19212409]
51. Zheng D. *Genome Biol* 2008;9:R105. [PubMed: 18598352]
52. Marques-Bonet T, Cheng Z, She X, Eichler EE, Navarro A. *BMC Genomics* 2008;9:384. [PubMed: 18699995]
53. She X, Cheng Z, Zollner S, Church DM, Eichler EE. *Nat Genet* 2008;40:909–14. [PubMed: 18500340]
54. Bailey JA, Liu G, Eichler EE. *Am J Hum Genet* 2003;73:823–34. [PubMed: 14505274]
55. Xing J, Witherspoon DJ, Ray DA, Batzer MA, Jorde LB. *Am J Phys Anthropol* 2007;(Suppl 45):2–19. [PubMed: 18046749]
56. Borges BN, Paiva TS, Harada ML. *Genetics and Molecular Research: GMR* 2008;7:663–78. [PubMed: 18752194]
57. Wise CA, Sraml M, Rubinsztein DC, Eastal S. *Mol Biol Evol* 1997;14:707–16. [PubMed: 9214743]
58. Rudd MK, Wray GA, Willard HF. *Genome Res* 2006;16:88–96. [PubMed: 16344556]
59. Alkan C, Ventura M, Archidiacono N, Rocchi M, Sahinalp SC, Eichler EE. *PLoS Comput Biol* 2007;3:1807–18. [PubMed: 17907796]
60. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. *PLoS Genet* 2007;3:e63. [PubMed: 17447845]

61. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Nat Genet 2008;40:1253–60. [PubMed: 18776909]
62. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. Nat Genet 2008;40:1166–74. [PubMed: 18776908]
63. McCarroll SA, Altshuler DM. Nat Genet 2007;39:S37–42. [PubMed: 17597780]
64. Lee C, Iafrate AJ, Brothman AR. Nat Genet 2007;39:S48–54. [PubMed: 17597782]
65. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Science 2007;316:445–9. [PubMed: 17363630]
66. Cahan P, Li Y, Izumi M, Graubert TA. Nat Genet. 2009
67. Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A. Nat Genet. 2009
68. Conrad DF, Hurler ME. Nat Genet 2007;39:S30–6. [PubMed: 17597779]
69. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. Nature 2008;451:998–1003. [PubMed: 18288195]

Figure 1.

An overview of array CGH technology. Total genomic DNA is sheared or enzymatically-digested, differentially-labeled and then co-hybridized to the slide containing hundreds of thousands (or millions) of oligonucleotide probes

Figure 2.

The distribution of Agilent 1M probes (Ag CGH 1X1M) at a known CNV region in human chromosome 22. The comparative nonhuman primate net tracks (marmoset, rhesus macaque, orangutan and chimpanzee) can reveal the extent of chromosome and sequence divergence from the human genome. Different colors on the net track indicate different chromosomes and gaps indicate sequence mismatches.

Figure 3.

A deletion in chimpanzee DNA sample relative to a pooled human reference visualized at a 396 kb region that starts at nucleotide position 96,259,545 and goes to nucleotide position 96,656,210 on human chromosome 10. Comparison of this region with the corresponding reference chimpanzee genome suggests that the orthologous sequences are indeed deleted in “Clint”

Figure 4.

Examples of CNV calls on the same 105K Agilent “human” array against the same human reference individual. The snapshot shows distribution of log2 ratios within a 2.65 Mb region on chromosome 2 (201,801,893-204,453,145) (based on human genome) with (a) a Chimpanzee sample and (b) an owl monkey sample. The snapshot was taken from DNA Analytics software (Agilent, Inc.).

Figure 5.

An example of genomic complexity within a CNV call on an Agilent 1M array (human sample vs. human reference). There is an additional copy number gain within the larger region of copy number gain. This snapshot was taken from DNA Analytics software (Agilent, Inc.).

Figure 6. Distribution of log2 ratios among individuals. Estimating the copy number state of (a) bi-allelic CNVs are easier than that of (b) multi-allelic CNVs

Figure 7.
An example for PCR based copy number validation visualized on 1% agarose gel.