Submicroscopic structural variations: A *de novo* tool for molecular anthropogenetics

Vipin Gupta[1,*], Rajesh Khadgawat[2], K. N. Saraswathy[1] and M. P. Sachdeva[1]

[1]Biochemical and Molecular Anthropology Laboratory, Department of Anthropology, University of Delhi, Delhi 110 007, India
[2]Department of Endocrinology and Metabolism, All India Institute of Medical Sciences, Ansari Nagar, New Delhi 110 029, India

In the quest for better genomic coverage and the need for a complete spectrum of genetic variability of complex human phenotypes, the role of larger variations, i.e. copy number variations (CNVs) excites medical geneticists and molecular anthropologists. CNVs are the technological 'missing link' filling the gap between the limits of sequence variation detection and traditional cytogenetic variation analysis. These variations are generally termed as structural variations, which includes microscopic or submicroscopic segments of DNA larger than 1 kb in size. Molecular anthropogeneticists must utilize this new tool for studying population structure defined in terms of genome diversity and hence must contribute with their efforts in the exploitation of genetic architecture of complex disorders.

Keywords: Copy number variation, polymorphisms, repeats, structural variations.

INTRIGUING submicroscopic variations are the technological 'missing link' filling the gap between the limits of sequence variation detection and traditional cytogenetic variation analysis. On the continuum of different variations in the human genome, one end comprises of sequence variations (like single nucleotide polymorphism) and the other end comprises of whole chromosomal variations (like aneuploidy, aneusomy, etc.). Between these two extremes of variation in the human genome lie the structural variations (a blanket term) which encompass cytogenetically visible and submicroscopic variations of the human genome. Cytogenetically visible structural variations are chromosomal deletions/insertions, inter/intra chromosomal translocations, etc., whereas submicroscopic structural variations includes copy number variations (CNVs), segmental duplications, submicroscopic inversions/translocations, etc. Therefore, CNVs are submicroscopic structural variations, which include submicroscopic segments of DNA larger than 1 kb, undertaken arbitrarily to accommodate the significant gap between smaller and larger variations[1]. Such submicroscopic variations typically reflect the unstable genomic regions having rearrangement of DNA due to regional genomic architecture and give rise to the concept of genomic disorders, which incorporate both inter- and intra-chromosomal rearrangements[2]. Human genomic variations occur on multiple levels, as the continuum begins from single nucleotide polymorphisms (SNPs) and ends at larger events involving contiguous blocks of DNA sequence that vary in copy number between individuals, and the efficacy to detect copy number variations on genome-wide scale has emerged only recently[3]. A full understanding of distribution of structural variations within species is necessary for the investigation of its medical and evolutionary impact on various populations[4]. Redon[5] estimated the population differentiation statistic by F_{ST} and its average value for autosomal CNVs was 0.11, similar to that observed for all autosomal phase I HapMap SNPs, i.e. 0.13. Recently, Nozawa *et al.*[6] showed the effect of genomic drift (random change of copy number during evolution) in generating intra- and inter-specific CNVs of sensory receptor genes. Nguyen *et al.*[7] argued that if large-scale DNA variations are beneficial, then they should be enriched in genes particularly involved in fighting infection and sensing the environment. They discovered such enrichments in the mouse genome, thus indicating the probable advantageous role of CNVs in human evolutionary history. These studies emphasize the urgent need of exploration of population genetics of CNVs, and the impact of evolutionary forces on these structural variations on different anthropologically well-defined ethnic groups across the globe.

Probable mechanism of presence of CNVs in human genome

Various probable mechanisms in the human genome are responsible for generating structural variations. For instance, retrotransposition of mobile elements like LINE elements, retroviruses, nonhomologous recombination, etc.[8]. Tuzun *et al.*[8] found conspicuous presence of structural variants near or in repetitive DNA regions of the human genome. Genomic disorders originate mostly from non-allelic homologous recombination (NAHR) between regions with specific low copy number repeats (LCR)[2].

Lack of technical standards in the detection of CNVs

The main approaches to identifying unbalanced structural variants are array-based analysis and quantitative, pri-

*For correspondence. (e-mail: udaiig@gmail.com)

mary PCR-based assays. Array-based comparative genomic hybridization (CGH) approaches provide the most robust methods for carrying out genome-wide scans to find novel CNVs with the help of a combination of bacterial artificial chromosomes (BACs) (starts from 50 kb) and long oligonucleotides (60–100 bp). To screen the target regions, the most robust assay is mainly based on PCR and the best established among them is real-time quantitative PCR. This works well for scoring individual deletions and duplications[9]. The progress regarding CNVs to date is largely due to the availability of numerous microarray platforms, which detect quantitative imbalances. And standard identification of variations requires a comparison either to a reference DNA source, a reference dataset or a reference genome sequence, which has implications for experimental design and interpretation of results. Unfortunately no standardized 'reference' control DNA has been adopted for laboratory experiments and in some cases 'pools' of samples or datasets are used to represent an averaged genome[1]. Most importantly, the validation of the findings of structural variants could only have been done with the help of an independent method[9]. Because no single approach identifies all types of structural variants, the standard terminology for acknowledging structural variation is also lacking. Researchers have also suggested that it may be better to use qualifiers for the term 'CNV' when discussing functional or clinical significance. They also suggested the terms 'pathogenic CNV', 'benign CNV' or CNV of unknown clinical significance[10]. To reduce the technical limitations and to have common standards of CNV research Scherer *et al.*[1] provide four broad guidelines: (i) Appropriately describing the origin of each sample, including all its other characteristics like age, sex, karyotypic status, phenotypic details, etc. (ii) Proper declaration of all aspects of experimental design and results. (iii) All studies should apply stringent quality control criteria to ensure an accurate empirical estimation of performance of the detection protocol used. (iv) All studies must thoroughly report all the properties of the structural variants, including sequence content, population frequency and its distribution. Current initiatives to discover and characterize structural variations are focused on simpler variations (>1 kb in size), because the detection of larger (submicroscopic) and more complex (may be more important) structural variants face various other confounding factors[1]; for instance, studies involving CNVs and linkage disequilibrium (LD) have typically excluded complex regions of the genome that are rich in duplications and prone to rearrangement[3].

Role of CNVs in complex disorders

Genomic variations have direct or indirect implications on phenotype and genotype relationships. Traditionally only a few reported genomic disorders (sporadic diseases) were known to be caused by *de novo* genomic structural alterations. Now their role in Mendelian disorders has been already appreciated, but not a single study is found in the context of complex disorders[11]. The roles of intermediate length deletion or duplication polymorphism contribute to common variation in healthy individuals[12]. In recent genomic surveys, the common copy number polymorphisms (CNPs) have been frequently found in genic regions, and have also been reported in several complex disorders. However, it has been hypothesized that common disorders are more susceptible to soft variations (variations in noncoding regions only altering the gene dosage)[11]. But the biomedical relevance of CNVs cannot be ignored on the basis of a few such studies. Clinical cytogeneticists also wish to clinically differentiate between CNVs that are pathogenic and those that are less likely to contribute to the affected phenotype with the help of available CGH technique[10]. Therefore, the role of such variations is an important aspect of both selection and susceptibility to disease[8].

The emerging CNV research is still in its discovery phase, i.e. generating a list of regions that contain CNVs, rather than focusing on association studies. Hence the underlying problems in CNV discovery and genotyping (for association) are different, due to different requirements behind the two hypotheses. Detection of the association of putative CNVs with clinical phenotypes also depends upon power of the study. This further relies upon the precise measurement of the allelic state (or genotype) of any CNV, which is still not well developed. Thus finding genotype–phenotype correlation with the help of CNVs poses both technical and analytical challenges. Researchers generally believe that CNVs will show lower correlation with clinical phenotypes in comparison to SNPs because of the greater challenge in measuring multibase, often multiallelic variants compared with single base, diallelic SNPs[11]. The common problem in reporting most of the CNV locations actually corresponds to the difficulty in locating potential CNV-containing regions (CNVRs). Moreover, even after the discovery of these regions, seldom it is known about the affected locus or gene within the CNVRs. Therefore, the exact locations of CNVs within the reported CNVRs are mandatory for researchers interested in validation of the reported CNVs in clinical samples from different populations[11].

The genetic dissection of complex disorders is more cumbersome because of the large genetic heterogeneity associated with it, which may also be influenced by *de novo* CNVs. For instance, Sebat *et al.*[13] tested the hypothesis that *de novo* CNVs (variants not present in their parents) are associated with autism spectrum disorders, and identified several *de novo* CNVs. The identified candidate regions were further validated by higher resolution comparative genomic hybridization, flourosence *in situ* hybridization, paternity testing, cytogenetics and micro-

sattelite genotyping, thus establishing their important role in genetic heterogeneity of complex disorders.

Practically, the development of assays for accurately typing CNVs in clinical samples has now become the most pressing need in CNV research. In view of the paucity of technical inputs in typing CNPs, a more discussed strategy might be to rely on more-easily-typed SNPs to serve as markers by LD for common variants throughout the genome to exploit LD between SNPs and CNPs. The success of this approach depends upon the strength and generality of LD between SNPs and CNPs. Assessing LD around CNPs requires accurate genotyping with dense SNP genotypes. Thus the extent of LD between the two classes of variants still remains unclear. Strong LD relation of very small common deletions and insertions with common SNPs suggests that although the mechanisms giving rise to them may be different, these polymorphisms share a similar evolutionary history[12]. In the light of recent ongoing genome-wide association studies, another practical suggestion is to integrate association studies for SNPs and CNVs. This step demands modifications in SNP genotyping assays to also incorporate CNVs (thus making hybrid arrays), but without much affecting the genome-wide coverage of SNPs, and improving the technical limitations of SNP assays specifically optimized for allelic discrimination (rather than CNVs)[11]. In contrast, Locke *et al.*[3] found only modest evidence of LD between CNPs and HapMap SNPs (there may be several possible reasons for such results). They cautioned researchers not to overinterpret their results, so that hope for the success of LD-based strategy still remains. Before using LD-based CNV mapping, it will be essential to determine whether rearrangement of genes on the human genome recurs in different genetic backgrounds. If mutational events occur too frequently, association studies based on LD of closely mapped SNP markers may not uncover an association with disease. In that case, the fine-scale structural mapping provides a better rationale for prioritizing regions for further studies[8]. Locke *et al.*[3] also observed reduced density of HapMap SNPs in regions of segmental duplication. This may result in the need for designing unique SNP assays in such regions. Undoubtedly, the overall impact of CNVs on human genomic variation is striking, and the inherent instability of these parts of the genome might also give rise to somatic CNVs that contribute to cancer progression and bipolar disorders[14]. It has been reiterated by researchers that as with SNP-based analysis, the importance of power, detection of population stratification, scrutinizing statistical thresholds, and *P*-value inflation also need careful scientific attention for CNV-based research[1,11].

De novo tool for molecular anthropogenetics

Functionally, biological anthropology is a comparative study of human variations among populations and/or communities in time and space. In simple words, studying human biological variations among populations within their socio-cultural milieu to have insightful learning is the only major goal of biological anthropology. Humans show great variation in phenotypic traits such as height, eye colour and susceptibility to disease, and this huge raw material (variation) for evolution is always an area of fascination for 'anthropogeneticists'. When anthropology faces its limits in terms of new tools for finding patterns of existing variation, a new platform for exploiting variations surfaces and challenges researchers towards the discovery of completely new vistas of anthropogenetic research. Recent reports suggest that intermediate and large-scale DNA structural variations are are an important source of genetic variation between individuals[12]. Thus the current job of molecular anthropologists is to find out the population-specific genetic variation of these submicroscopic variations. In view of the endogamous marriage pattern in India, it will be interesting to see the population structure in light of these structural variations among different ethnically well-defined population groups. Furthermore, the discovery of novel CNVs on disease-specific genic regions demands their validation in other populations, to have a better estimate of the allelic architecture of the concerned disease explained in terms of structural variations. Hence there is an urgent need to redesign the current platforms to study the genome, either to maximize detection of CNVs or minimize their interference with available methods to detect other forms of genomic variations[15].

A typical example of molecular anthropogenetic research is provided by Perry *et al.*[16], where they had compared the 355 CNVs between chimpanzees (*Pan troglodytes*) and humans and found the loci of ancestral segmental duplications, some of which may be unstable hotspots for the genesis of CNVs. Again, the perspective of molecular anthropology (i.e. incorporating cultural aspects of the population) can easily be elucidated with the example of copy number of salivary amylase gene (*AMY1*), which is found to be positively correlated with the salivary amylase protein level and that individuals from a population with high-starch diets (agricultural societies and hunter–gatherers in arid environment), have on an average more copy numbers of the *AMY1* gene[17].

The commercially available fixed marker sets (like Affymetrix, Illumina, etc.) have also begun including CNVs to have a more powerful combination of markers providing better genome-wide coverage in comparison to only SNP-based arrays. For instance, Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 probes for the detection of CNV. This particular array represents more genetic variation on a single array than any other product, providing maximum panel power and the highest physical coverage of the genome. These powerful arrays could be used to reveal the population

history of different ethnic groups of the world, as done by Hughes *et al.*[18] recently, using only the SNP-based array. Utility of such high-density CNV-based arrays in genome diversity studies would be a revolution. In countries like India, where the entire population could easily be divided into various large endogamous groups (who marry among themselves) on the basis of deeply integrated factors like caste and religion rather than only geography, survey of CNVs at genome level would be helpful in detecting ancestry informative markers. SNP-based databases are already available, but similar efforts in case of CNVs are yet to be realized. For instance, Wong *et al.*[19] performed a whole genome analysis of CNVs and identified 800 reasonably polymorphic autosomal segmental CNVs which appeared at a frequency of at least 3% and created baseline human genetic variation. They also suggest the role of CNVs in human phenotypic variation. In the Indian population, which reflects highly stratified ethnic groups, if we want to get optimum results, such databases (for CNVs) should be endogamous-population based. Baris *et al.*[20] tried to prove the diagnostic utility of array-based comparative genomic hybridization (aCGH) as an adjunct to chromosomal analysis tests in the evaluation of patients having chromosomal disorders. They also suggested that routine usage of aCGH in clinical settings will lead to better understanding of submicroscopic structural abreactions. Therefore, such a targeted aCGH could be helpful in detecting putative CNV variations among phenotypically normal individuals in various ethnic groups.

Thus, the potential clinical relevance of a CNV increases with the simultaneous increment in the number of genes within the region of genomic imbalance. Moreover, it has also been generally thought that duplication CNVs are better tolerated in the genome than deletion CNVs (have higher likelihood of being pathogenic)[10]. Now, because CNVs are a part of the contemporary population genetics discourse (On 6 August 2008, the total number of CNV entries reached 17,641 on 5672 loci in the Database of Genomic Variants) on genomic variation studies and their biological, health and clinical implications, hence molecular anthropogeneticists must profile CNVs in their ethnicity-based genetic epidemiological research, and design proposals with the help of available arrays of fixed marker sets for strengthening the pattern of genetic architecture of complex disorders.

1. Scherer, S. W. *et al.*, Challenges and standards in integrating surveys of structural variation. *Nature Genet.*, 2007, **39**, S7–S15.
2. Stankiewicz, P. and Lupski, J, R., Genomic architecture, rearrangements and genomic disorders. *Trends Genet.*, 2002, **18**, 74–82. s
3. Locke, D. P. *et al.*, Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, 2006, **79**, 275–290.
4. Conrad, D. F. and Hurles, M. E., The population genetics of structural variation. *Nature Genet.*, 2007, **39**, S30–S36.
5. Redon, R. *et al.*, Global variation in copy number in the human genome. *Nature*, 2007, **444**, 444–457.
6. Nozawa, M., Kawahara, Y. and Nei, M., Genomic drift and copy number variation of sensory receptor in humans. *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 20421–20426.
7. Nguyen, D., Webber, C. and Pointing, C. P., Bias of selection on human copy-number variants. *PLoS Genet.*, 2006, **2**, 198–207.
8. Tuzun, *et al.*, Fine-scale structural variation of the human genome. *Nature Genet.*, 2005, **37**, 727–732.
9. Feuk, L., Carson, A. R. and Scherer, S. W., Structural variation in the human genome. *Nature Rev. Genet.*, 2006, **7**, 85–97.
10. Lee, C., Iafrate, A. J. and Brothman, A. R., Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genet.*, 2007, **39**, S48–S54.
11. McCarroll, S. A. and Altshuler, D., Copy-number variation and association studies of human disease. *Nature Genet.*, 2007, **39**, S37–S41.
12. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. and Frazer, A. F., Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genet.*, 2006, **38**, 82–85.
13. Sebat, J. *et al.*, Strong association of *de novo* copy number mutations with autism. *Science*, 2007, **316**, 445–449.
14. Kehrer-Swatzki, H., What difference copy number variation makes. *BioEssays*, 2007, **29**, 311–313.
15. Hegle, R. A., Copy number variations add new layer of complexity in the human genome. *CMAJ*, 2007, **176**, 441–442.
16. Perry, G. H. *et al.*, Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl. Acad. Sci. USA*, 2006, **103**, 8006–8011.
17. Perry, G. H. *et al.*, Diet and the evolution of human amylase gene copy number variation. *Nature Genet.*, 2007 (September online publication).
18. Hughes, Welch, R., Puri, V., Mathews, C., Haque, K., Chanock, S. J. and Yeager, M., Genome-wide SNP typing reveals signatures of population history. *Genomics*, 2008.
19. Wong, K. K. *et al.*, A comprehensive analysis of common copy-number variations in human genome. *Am. J. Hum. Genet.*, 2007, **80**, 91–104.
20. Baris, H. N., Tan, W., Kimonis, V. E. and Irons, M. B., Diagnostic utility of array-based comparative genomic hybridization in a clinical setting. *Am. J. Med. Genet. Part A*, 2007, **143**, 2523–2533.

Received 2 February 2008; revised accepted 9 July 2008