**RESEARCH NOTE**

# SNPs in genes with copy number variation: A question of specificity

MAINAK SENGUPTA[1], ANANYA RAY[1,2], MOUMITA CHAKI[1], MAHUA MAULIK[1,3] and KUNAL RAY[1,*]

[1]*Molecular and Human Genetics Division, Indian Institute of Chemical Biology, Kolkata 700 032, India*
[2]*Current address: University of Rochester, Rochester, New York 14627, USA*
[3]*Current address: Centre for Neuroscience, University of Alberta T6G 2S2, Canada*

## Introduction

The specificity of single nucleotide polymorphisms (SNPs) is likely to be compromised with most of the current PCR-based methods used to genotype a target locus in the presence of a highly homologous duplicated region. Such a lack of locus specificity could inflate the heterozygosity of the SNPs. We reasoned that public database for SNPs might be influenced by false allele calls, specifically in genes with copy number variation (CNV). Therefore, we compared the fraction of SNPs with high heterozygosity values ($\geq$ 0.4) in NCBI dbSNP for genes with and without CNVs. Our observation highlights the challenges of selecting SNPs in genes with CNV for usage in complex biological studies.

SNPs have become a key tool in investigating how genes interplay in complex diseases. Based on the common disease – common variant (CD-CV) hypothesis, alleles with high heterozygosity are normally preferred. However, in duplicated regions of the genome with a high level of homology, a lack of locus specificity of the amplicon used to genotype nucleotide variants could inflate the density of SNPs (Bailey *et al*. 2002). That is, the nucleotides that are present in equivalent locations of the duplicated regions, but vary between the two loci, would appear to have a heterozygous genotype by inadvertent coamplification of the duplicated region(s). This single base variation in the equivalent position of the duplicon is known as a paralogous sequence variant (PSV), or cismorphism, and the presence of such a PSV would always yield a heterozygous reading when genotyped. Also, if the nucleotide at an equivalent position of a duplicon is same as that of the actual genomic region, one must take appropriate measures to determine whether the polymorphism resides in the gene or a duplicated region. Observations would be even further complicated if the base is polymorphic at more than one site, as in case of multi-site variants (MSVs) (Fredman *et al*. 2004). It is likely that any of the above cases could be misread as SNPs, and thereby elevate the ostensible heterozygosity of the probing base of the target gene. The recent discovery of CNV in the genome (Redon *et al*. 2006) has increased the complexity of the matter as they range from thousands to millions of DNA bases, and differ from one individual to another. Further, in most cases, the extent of variation is unknown. The strategies used to locate spurious SNPs and circumvent related problems in regions of the genome with well characterized duplicons cannot be applied to newly described CNVs that are not yet well defined in terms of the level of variation.

We hypothesized that to the unsuspecting investigator the heterozygosity values would appear higher in genomic regions with duplication than in those without it, resulting from false assignment due to lack of locus specificity in most of the currently used PCR-based methods. In this context, we were interested to determine whether SNPs in the genes recently described to have CNV, with potentially duplicated regions, have high heterozygosity values depicted in NCBI dbSNP.

## Materials and methods

We selected two sets of genes from dbSNP; the first known to have CNVs in the form of duplication (*FCGR3B*, the first 6 exons of *DISC1, CRYBB2, CRYBB3* and *PSORS1*) (Redon *et al*. 2006), or segmental duplication (*GH1, CSH1* and *PSORS1*), and a control set for which no such duplicons have been reported (*P53, CDC20, MDM2, CDKN1A, IL1A, MYOC*) (database of genomic variants: http://projects.tcag.ca/variation/). For each gene, we pooled all the SNPs for which heterozygosity values were reported and calculated the fraction of SNPs with high heterozygosity ($\geq$ 0.4) and

*For correspondence. E-mail: kunalray@gmail.com.

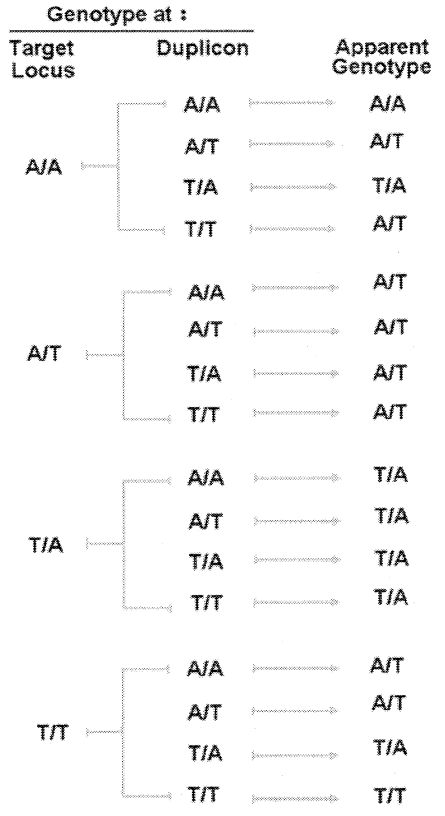
**Keywords.** copy number variation (CNV); single nucleotide polymorphism (SNP); heterozygosity.

compared them for genes with and without CNVs and/or segmental duplication.

## Results and discussion

For easier understanding of the underlying problem in genes with CNV, we first calculated the actual and apparent genotypes at a hypothetical locus with a defined duplicated region (figure 1a). We show that the probability of obtaining a heterozygous genotype increases from 50% to 87.5%, if both the alleles exist in equal proportion (figure 1b).

a




b

| Genotype at Target Locus: | Actual | Apparent |
|---|---|---|
| A,A | 25% | 6.25% |
| A,T | 50% | 87.5% |
| T,T | 25% | 6.25% |

**Figure 1.** An estimation of true versus observed genotypes of SNPs in genes with duplicated regions. For (a) illustrates how the presence of a duplicon affects the observed genotype at a target locus, and (b) Shows the increment in the observed fraction of heterozygotes as compared to the actual fraction. The percentage of each genotype was calculated assuming that A and T exist in equal frequencies, i.e. the allelic frequency of A = 0.5, and that of T = 0.5.

It is note worthy that in genes with CNV, such estimation is not easily attainable due to potential variation in the number of copies of the gene between individuals and different population groups. However, the overall apparent heterozygosity of the locus would be expected to be higher (due to false allele calls) compared to genes without duplicated regions. In this context, we took two groups of genes (as described in materials and methods): (i) genes with segmental duplication or reported to have CNV, and (ii) genes with a unique sequence in the genome (or with low level of homology, if any, with other genomic region). Our results showed that the relative level of heterozygosity of SNPs for the group of genes reported to have CNVs is consistently higher than in the control group (i.e., 0.38–0.75 versus 0.08–0.21) as shown in figure 2, and the difference is estimated to be statistically significant ($P$-value $< 0.0075$; unpaired t–test, Mann–Whitney and Kolmogorov–Smirnov test).

Therefore, it seems reasonable that PSVs and MSVs may in fact be contaminating the SNP profiles reported in public databases for genes with CNVs or segmental duplication.

Such a difference in the fraction of SNPs with high heterozygosity was not apparent from the highly homologous region between *NAT1* and *NAT2*. The duplicated region (~2.5 kb) being within the limits of PCR amplification and thereby the feasibility to design locus-specific amplicons is a plausible reason for this observation.

Another interesting case was the alpha globin gene family, which is known to have a high level of homology between its member genes (*HBA1, HBA2, HBZ, HBQ1* and *HBM*). SNPs described at these loci are likely to be affected by genotyping error due to lack of locus specificity of the targeted amplicon. We noticed that the previous NCBI built, Build 126 reported a large number of SNPs in regions with potential PSVs and MSVs (see table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/), which have been removed from the reference assembly of Build 127 and are reported with the header 'mapped unambiguously in non-reference assembly only'. However, Build 127 still contains SNPs in *PSORS1* and *GH1* that are likely to be PSV and MSV contaminants (see table 2 (a&b) in electronic supplementary material).

Our investigation demonstrates that in genomic regions containing CNVs and/or segmental duplication, some of the highly heterozygous nucleotide variants (SNPs) stored in publicly available databases are likely to be erroneously recorded PSVs and MSVs. Explicit knowledge regarding variation in copy number of the gene of interest in the study population is a prerequisite for accurate allelic information at the target SNP. These factors must be considered prior to the usage of specific SNPs as 'tools' to assess the contribution of genes in complex diseases, drug response, susceptibility of hosts to various pathogens and other related studies.

**Acknowledgements**

We thank Dr Arijit Mukhopadhyay for helpful discussion at the initial phase of the study. The study has been partially supported by the Council of Scientific and Industrial Research (CSIR), Govt. of

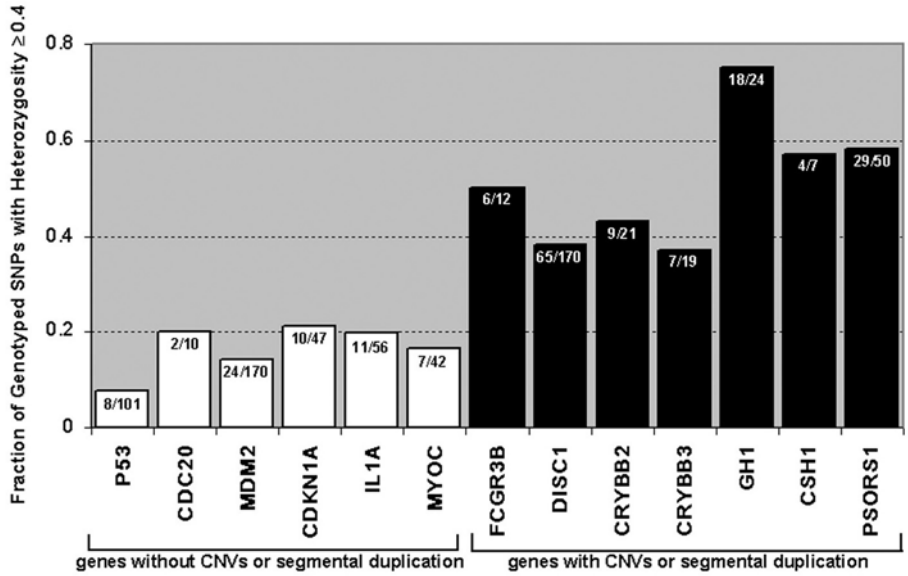



**Figure 2.** Genes with CNV and/or segmental duplication have a significantly higher number of SNPs with high heterozygosity in NCBI dbSNP. The fraction of highly heterozygous (≥ 0.4) SNPs reported in NCBI dbSNP (Build 127) were found to be significantly higher in genes with CNVs and/or segmental duplication (denoted by black bars) than in single copy genes (white bars). The actual ratio of the SNPs (with high and low heterozygosity) retrieved from the database for each gene is shown at the top of the corresponding bar. The *PSORS1* gene corresponds to a reported segmental duplication and a CNV region.

India (Grant No: CMM 0016). MS and MC are supported by predoctoral fellowships from University Grant Commission (UGC).

## References

Bailey J. A., Gu Z., Clark R. A., Reinert K., Samonte R. V., Schwartz S. *et al*. 2002 Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.

Fredman D., White S. J., Potter S., Eichler E. E., Den Dunnen J. T. and Brookes A. J. 2004 Complex SNP-related sequence variation in segmental genome duplications. *Nat. Genet*. **36**, 861–866.

Redon R., Ishikawa S., Fitch K. R., Feuk L., Perry G. H., Andrews T. D. *et al*. 2006 Global variation in copy number in the human genome. *Nature* **444**, 444–454.

Received 20 November 2007; accepted 24 January 2008
Published on the Web: 2 April 2008