

# Copy-number polymorphisms: mining the tip of an iceberg

Patrick G. Buckley<sup>\*</sup>, Kiran K. Mantripragada<sup>\*</sup>, Arkadiusz Piotrowski, Teresita Diaz de Ståhl and Jan P. Dumanski

Department of Genetics and Pathology, Rudbeck laboratory, Uppsala University, 751 85 Uppsala, Sweden

**Copy-number polymorphisms (CNPs) represent a greatly underestimated aspect of human genetic variation. Recently, two landmark studies reported genome-wide analyses of CNPs in normal individuals and represent the beginning of an understanding of this type of large-scale variation. Future array-CGH-based CNP analyses should include standard criteria on a common microarray platform. It is only when parallel analyses of CNPs and SNPs are performed in an integrated format that we will obtain a global picture of our genetic diversity.**

## Introduction

The study of human genetic variation at the DNA level constitutes a major challenge and has received considerable attention in the post-genomic era. The dominating type of variation explored so far in the genome has been single nucleotide polymorphisms (SNPs), overshadowing the issue of copy-number polymorphisms (CNPs) (gains and deletions) [1]. The current approach to study genetic variation can be viewed as biased, in the sense that the identification of genome-wide large-scale CNPs is virtually untouched compared with detailed analyses of millions of SNPs. We believe that analysis of SNPs and CNPs are necessary to obtain a more complete picture of our genetic diversity.

The presence of a limited number of the best-studied form of DNA copy-number variation (indels) was previously observed in the human genome, and several studies have ascertained their importance in health and disease [2–8]. For example, in a study of ovarian cancer cell lines, Lin *et al.*, identified a 276-bp region of chromosome 22q13 that was deleted not only in 47% of ovarian cancer cell lines but also in 18% of constitutional DNA samples from healthy individuals [6]. Another study reported a 102-bp homozygous deletion on chromosome 8p12–21 in biliary tumors and pancreatic tumors (and cell lines) as well as in normal individuals [9]. Both of these studies concluded that the identified deletions might represent normal human genetic variation rather than cancer-associated aberrations. Interestingly, some indels provide a protective effect against disease in ‘normal’ (or unaffected) individuals. For example, the 32-bp deletion

polymorphism of the C-C chemokine receptor 5 gene (*CCR5*) confers a reduced susceptibility to HIV-1 infection in homozygous individuals [10]. This demonstrates that the hidden functionality of such normal variation only becomes apparent when challenged by environmental factors.

The genome-wide detection of CNPs has been complicated because of the lack of high-resolution and high-throughput techniques. A fundamental step towards identifying such variation was the development of microarray-based comparative genomic hybridization (array-CGH) [11,12]. This method is based on the assessment of fluorescence ratios between differentially labeled test and reference DNA, hybridized to a microarray [13,14]. Altered fluorescence ratios are therefore indicative of DNA copy-number imbalance (loss or gain) in the test versus the reference genome. Although many array-CGH studies that focus on tumor-associated gains or deletions (indicating the presence of activated oncogenes or inactivated tumor-suppressor genes) have been performed [15–18], there are few studies addressing this type of variation in normal individuals.

## Genome-wide array-based detection of CNPs

Recently, two landmark studies have reported the presence of CNPs in humans using different genome wide array-CGH based techniques [19,20]. Iafrate *et al.* used commercially available bacterial artificial chromosome (BAC) arrays, whereas Sebat *et al.* developed and applied a custom-made oligonucleotide array for the assessment of copy-number variation. A brief comparison of these studies is presented in Table 1. Both studies convincingly demonstrate the presence of genomic imbalances among normal individuals, which overlap with genes and often coincide with segmental duplications in the genome and can contribute to phenotypic variation and disease susceptibility. In these reports, Sebat *et al.* and Iafrate *et al.* identified 76 and 255 loci, respectively, that display copy-number variation in the human genome. One of the disadvantages of both studies is the limited number of samples used to assess DNA copy-number variation. One would expect at least 100 individuals to be analyzed to assess the frequency of CNPs using similar standards as applied for SNPs (i.e. a change detected in <1% of analyzed samples is considered a mutation, rather than a polymorphism). However, this suggested number might need to be revised, because the overall frequencies of private [i.e. unique to single individual (or kindred) or

Corresponding author: Dumanski, J.P. (jan.dumanski@genpat.uu.se).

<sup>\*</sup> Both of these authors contributed equally.

Available online 26 April 2005

**Table 1. Comparison between two recent reports on global genome analysis of DNA copy-number polymorphisms (CNPs)**

Aspects of the studies	Sebat <i>et al.</i> [20]	Iafrate <i>et al.</i> [19]
Array construction and study design	70mer oligonucleotides, in combination with ROMA <sup>a</sup>	BAC clones; conventional array-CGH
Array coverage	One measurement point every 35 kb <sup>b</sup>	One measurement point every 1Mb <sup>b</sup>
Individuals tested	20 phenotypically normal	39 phenotypically normal; 16 with previously identified chromosomal abnormalities
Number of different ethnic populations	9	5
CNP loci detected	76	255
Average frequency of CNP per individual	11	12.4
Most common gain or loss	A gain (90%) at 14q11; a loss (90%) at 15q11	A gain (31%) at 14q12; a loss (38%) at 13q21.1
Major disadvantage	Relies on restriction digest representation of the human genome	Limited resolution

<sup>a</sup>Abbreviation; ROMA, representational oligonucleotide microarray analysis.

<sup>b</sup>The average resolution of the two studies is difficult to compare on the basis of the presented data.

rare] versus common CNPs are currently unknown. Another striking aspect of these studies is that neither has observed variation at each others most frequent polymorphic sites (Table 1). For example, the most common CNPs detected by Sebat *et al.* that were polymorphic in 90% of studied cases (14q11 gain of 201 kb and 15q11 loss of 1.57 Mb) were not detected by Iafrate *et al.* Therefore, the ability or failure to detect such variation might largely rely on the genome coverage and the nature of the DNA used for microarray construction. Interestingly, Sebat *et al.*, who used tiled oligos, claim to have higher resolution than the BAC-array approach used by Iafrate *et al.*, but detected three times less polymorphic genomic loci. This can probably be explained by the differences in experimental design (e.g. the representational amplification step used by Sebat *et al.*) or in the selection bias of analyzed samples.

There is a need for a standard terminology when describing DNA copy-number variation. Iafrate *et al.* used the term large-scale variation (LSV), whereas Sebat *et al.* used CNP (copy-number polymorphism) to define copy-number changes that they detected in the genome. Indel (insertion or deletion) is a well-established and accepted term that refers usually to deletion or insertion variation of a limited number of bases. We endorse the term CNP, which could be used to describe any DNA copy-number variation (di-allelic or multi-allelic) that is >200 bp. This category of genetic polymorphism should not include the variation based on multiallelic tandem repeats [i.e. minisatellites or variable number of tandem repeats (VNTRs) and microsatellites also called short tandem repeats (STRs)].

### Initial analysis of chromosome 22 CNPs

Our knowledge of the extent of CNPs in the human genome is likely to increase considerably with advancements in technologies. The current resolution of array-CGH using total genomic DNA enables analysis at the exon level [21]. Our group is focused on the development and application of genomic microarrays for the detection of copy-number alterations in patient cohorts and in normal individuals. We have developed a genomic clone-based microarray covering human chromosome 22 with an average resolution of 75 kb [22]. After identification of loci prone to DNA copy-number alterations using this array, a more detailed analysis is performed using a PCR-based repeat-free strategy for the construction of arrays with an average resolution of 2 kb [23]. The

array-CGH data obtained so far suggest that there exists a wide range of copy-number variation on chromosome 22q. From our experience, using the PCR-based strategy, the higher the resolution of analysis the more CNPs detected. For example, deletions and/or amplifications have been observed in the Cat-Eye syndrome, immunoglobulin lambda locus, the *LARGE* gene (encoding like-glycosyltransferase), glutathione S-transferase  $\theta$  1 (*GSTT1*) and several other loci on 22q, ranging from 200 bp to 3 Mb (C. de Bustos, *et al.*, unpublished; T. Diaz de Stahl *et al.*, unpublished; P.G. Buckley *et al.*, unpublished). When we were able to study families, we observed that these CNPs show transmission through generations. A large proportion of the CNPs detected also coincide with segmental duplications, which point to their importance in the mechanism generating this variation. Thus, we believe that the results from Iafrate *et al.* and Sebat *et al.* represent the 'tip of the CNP iceberg'. However, it is currently difficult to estimate its size. Based on our preliminary work from chromosome 22q (~1% of the genome), it seems inappropriate to extrapolate the frequency of observed CNPs to the rest of the genome, when we consider that chromosome 22 is rich in segmental duplications.

In contrast to SNPs (which are usually binary +/–), the analysis of CNPs is considerably more difficult owing to the complexity of variation and the expense involved in their detailed assessment. It is not only the presence or absence of the CNP that matters, but the type (i.e. zero, one, two, three copies) and physical extent of duplication or deletion must also be considered. We can illustrate this by our experience with the known CNP of the *GSTT1* locus [24] on 22q, using the repeat-free and non-redundant array-CGH platform mentioned previously [23]. Based on the analysis of 100 normal individuals, we observed two, one and zero copies of this locus, which reflects one level of complexity. An additional level is represented by the possible differences in the size of such polymorphic deletions or gains. This can possibly affect other genes in the vicinity of CNPs within gene-rich segments of 22q. However, these size differences are difficult to estimate using available methodologies because *GSTT1* is flanked by segmental duplications and has a high content of common repeats.

### Concluding remarks

Currently, the approaches for analysis of copy-number variation include the use of arrays of genomic clones

(BACs, PACs and cosmids), cDNA clones, PCR products and oligonucleotides. The comparability of the results generated from these different platforms is currently the major obstacle in the field and contributes to confusion in data interpretation between different reports. Furthermore, there is a lack of common uniform criteria for the quality assessment of published array-CGH data. A standard similar to the 'minimum information about microarray experiment' (MIAME; <http://www.mged.org>) expression array criteria should be introduced. The recently reported 32K human BAC array [25] could serve as an initial standard platform for genome-wide large-scale analysis of CNPs. We advocate that to study disease-associated copy-number variation accurately, a baseline of CNP frequency should be established (similar to the approach used for SNPs) by analyzing a large sample group derived from multi-ethnic backgrounds using this resource. The data generated should be processed using standardized criteria and compiled in a publicly available database. The identified polymorphic sites should be followed up with detailed analyses using higher resolution arrays. However, even in the case where the majority of CNPs and SNPs are characterized, the challenge of understanding the phenotypic effects of both types of variation still remains. It is only when the majority of CNPs and SNPs in the human genome are characterized and their phenotypic effects understood that we can begin to understand our genetic differences.

#### Acknowledgements

This work was supported by grants from the U.S. Army Medical Research and Materiel Command award no. W81XWH-04-1-0269, the Swedish Cancer Foundation, the Swedish Research Council and Uppsala University. We thank Ian Dunham and Carl Bruder for critical comments.

#### References

- 1 Siniscalco, M. *et al.* (2000) A plea to search for deletion polymorphism through genome scans in populations. *Trends Genet.* 16, 435–437
- 2 Barber, J.C. *et al.* (1998) Duplication of 8p23.1: a cytogenetic anomaly with no established clinical significance. *J. Med. Genet.* 35, 491–496
- 3 Ghanem, N. *et al.* (1988) Polymorphism of MHC class III genes: definition of restriction fragment linkage groups and evidence for frequent deletions and duplications. *Hum. Genet.* 79, 209–218
- 4 Groot, P.C. *et al.* (1991) Interpretation of polymorphic DNA patterns in the human alpha-amylase multigene family. *Genomics* 10, 779–785
- 5 Engelen, J.J. *et al.* (2000) Duplication of chromosome region 8p23.1 → p23.3: a benign variant? *Am. J. Med. Genet.* 91, 18–21
- 6 Lin, H. *et al.* (2000) A frequent deletion polymorphism on chromosome 22q13 identified by representational difference analysis of ovarian cancer. *Genomics* 69, 391–394
- 7 Buckland, P.R. (2003) Polymorphically duplicated genes: their relevance to phenotypic variation in humans. *Ann. Med.* 35, 308–315
- 8 Ciccoira, M. *et al.* (2004) Effects of ACE gene insertion/deletion polymorphism on response to spironolactone in patients with chronic heart failure. *Am. J. Med.* 116, 657–661
- 9 Ryu, B. *et al.* (2001) Frequent germline deletion polymorphism of chromosomal region 8p12-p21 identified as a recurrent homozygous deletion in human tumors. *Genomics* 72, 108–112
- 10 Carrington, M. *et al.* (1999) Genetics of HIV-1 infection: chemokine receptor CCR5 polymorphism and its consequences. *Hum. Mol. Genet.* 8, 1939–1945
- 11 Solinas-Toldo, S. *et al.* (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* 20, 399–407
- 12 Pinkel, D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20, 207–211
- 13 Albertson, D.G. and Pinkel, D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* 12, R145–R152
- 14 Mantripragada, K.K. *et al.* (2004) Genomic microarrays in the spotlight. *Trends Genet.* 20, 87–94
- 15 Bruder, C.E. *et al.* (2001) High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum. Mol. Genet.* 10, 271–282
- 16 Wilhelm, M. *et al.* (2002) Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Res.* 62, 957–960
- 17 Redon, R. *et al.* (2002) Amplicon mapping and transcriptional analysis pinpoint cyclin L as a candidate oncogene in head and neck cancer. *Cancer Res.* 62, 6211–6217
- 18 Wessendorf, S. *et al.* (2003) Hidden gene amplifications in aggressive B-cell non-Hodgkin lymphomas detected by microarray-based comparative genomic hybridization. *Oncogene* 22, 1425–1429
- 19 Iafrate, A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949–951
- 20 Sebat, J. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science* 305, 525–528
- 21 Dhami, P. *et al.* Exon Array-CGH: detection of copy number changes of individual exons in human disease Genes. *Am. J. Hum. Genet.* (in press)
- 22 Buckley, P.G. *et al.* (2002) A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications. *Hum. Mol. Genet.* 11, 3221–3229
- 23 Mantripragada, K.K. *et al.* (2003) Development of *NF2* gene specific, strictly sequence defined diagnostic microarray for deletion detection. *J. Mol. Med.* 81, 443–451
- 24 Landi, S. (2000) Mammalian class theta GST and differential susceptibility to carcinogens: a review. *Mutat. Res.* 463, 247–283
- 25 Ishkanian, A.S. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* 36, 299–303