MINI REVIEW

Genomics & Informatics Vol. 6(1) 1-7, March 2008

Copy Number Variations in the Human Genome: Potential Source for Individual Diversity and Disease Association Studies

Tae-Min Kim[1], Seon-Hee Yim[2] and Yeun-Jun Chung[1,2]*

[1]Department of Microbiology, [2]Integrated Research Center for Genome Polymorphism, The Catholic University of Korea, Seoul 137-701, Korea

Abstract

The widespread presence of large-scale genomic variations, termed copy number variation (CNVs), has been recently recognized in phenotypically normal individuals. Judging by the growing number of reports on CNVs, it is now evident that these variants contribute significantly to genetic diversity in the human genome. Like single nucleotide polymorphisms (SNPs), CNVs are expected to serve as potential biomarkers for disease susceptibility or drug responses. However, the technical and practical concerns still remain to be tackled. In this review, we examine the current status of CNV DBs and research, including the ongoing efforts of CNV screening in the human genome. We also discuss the characteristics of platforms that are available at the moment and suggest the potential of CNVs in clinical research and application.

Keywords: array-CGH, Copy number variation (CNV), Genome-wide association study (GWAS)

Introduction

Traditionally, large-scale genomic variants that are visible in conventional karyotyping have been thought to be associated with early-onset, highly penetrant genetic disorders, while they are incompatible in normal, disease-free individuals (Lupski, 1998; Stankiewicz and Lupski, 2002). The construction of the 'reference genome' by the human genome sequencing project is based on the belief that human genome sequences are virtually identical, even in different individuals, except for well-known single nucleotide polymorphisms (SNP) or size-variants of tandem repeats such as mini- or micro-satellites (variable number of tandem repeats or VNTR) (Przeworski *et al.*, 2000). This traditional concept has been recently challenged by the discovery that large structural variations are more prevalent than previously presumed (Check, 2005). Using high-resolution whole-genome scanning technologies such as array-based comparative genomic hybridization (array-CGH), two groups of pioneering scientists have identified widespread copy number variations (CNVs) in apparently healthy, normal individuals (Iafrate *et al.,* 2004; Sebat *et al.*, 2004). It proposes that our genome is more diverse than has ever been recognized, and subsequent studies have identified up to 11,000 CNVs across the whole genome (Tuzun *et al.*, 2005; Hinds *et al.*, 2006; Mills *et al.*, 2006; McCarroll *et al.*, 2006; Conrad *et al.*, 2006; Sharp *et al.*, 2005; Wong *et al.*, 2007; de Smith *et al.*, 2007).

Although the current understanding of CNVs is still limited for practical use and technical challenges still remain to be tackled, recent studies already have demonstrated the potential association of CNVs with various diseases, suggesting plausible functional significances and highlighting the promising utility of CNVs.

The current coverage of CNVs in the human genome already has exceeded that of SNPs (approximately 600 Mb comprising ～12% of human genome) and is still increasing (Cooper *et al.*, 2007). These large-scale structural variants, in addition to SNPs, will serve as powerful sources to help our understanding of human genetic variation and of differences in disease susceptibility for various diseases. This paper reviews the current knowledge and future perspectives of CNVs.

The definition of CNV

Structural variations that involve large DNA segments can take various forms, such as duplication, deletion, insertion, inversion, and translocation. Among them, DNA copy number variations larger than 1 kb are collectively termed CNVs. Fig. 1 illustrates the concept of CNV. Although the CNV can include large, microscopically visible genomic variations, it generally indicates a submicroscopic structural variation that is hardly detectable by conventional karyotyping (3～5 Mb) (Freeman *et al.*, 2006). Smaller variations such as small insertional- deletion (indel) polymorphisms are not included in CNVs, while they comprise another large collection of over

*Corresponding author: E-mail yejun@catholic.ac.kr
Tel +82-2-590-1214, Fax +82-2-596-8969

Accepted 11 March 2008

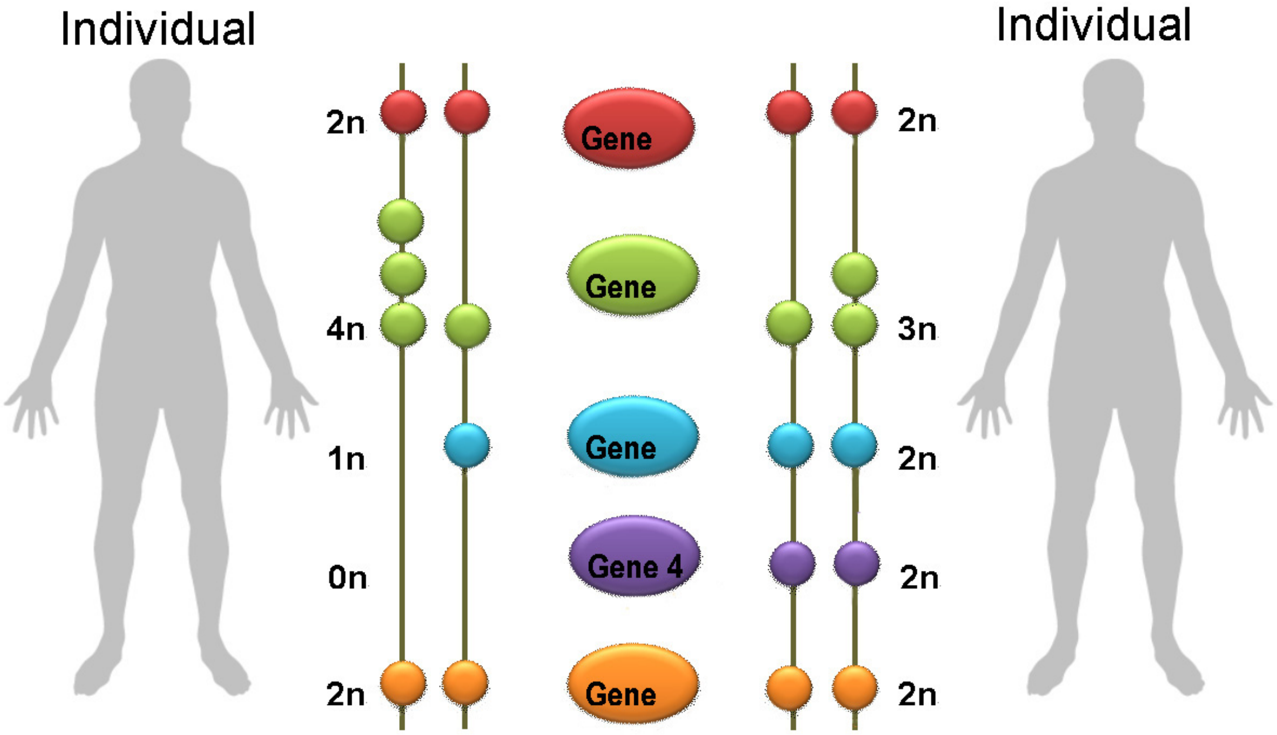

Fig. 1. Concept of inter-individual copy number variations. Most humans have two copies of DNA/gene (diploid, 2n), each originating from the maternal (n) and paternal (n) genome. However, some of the DNA regions/genes are not diploid, showing variable numbers of copies. These variable copies are mainly inherited but partly can be gained *de novo*. This phenomenon is termed copy number variation (CNV). Such CNVs may affect gene expressions through dosage imbalance of genes.

400,000 variants in the human genome (Mills *et al.*, 2006), and neither is the insertional polymorphism of mobile elements such as Alus or L1 elements considered a CNV.

At the beginning stages of CNV discovery, a number of terms were proposed to define them *e.g.*, large-scale copy number variants (LCV) (Iafrate *et al.*, 2004), copy number polymorphism (CNP) (Sebat *et al.*, 2004), and intermediate-sized variants (ISV) (Tuzun *et al.*, 2005). The current definition of CNV is also operational and can be modified with the advance of scanning resolution and coverage, and availability of allele frequency in a determined population.

The identification of CNVs using different platforms

Various scanning platforms and quality control methods have been used to identify CNV calls. Because the choice of platforms has a great effect on the results, it is worth reviewing the characteristics of platforms to improve the understanding of CNVs.

The presence of CNVs in normal individuals was reported for the first time in 2004 independently by two groups led by Lee C. and Wigler M. (Iafrate *et al.*, 2004; Sebat *et al.*, 2004). Both studies used two-dye array-CGH techniques that used clones of bacterial artificial chromosomes (BAC) or oligonucleotides (representational oligonucleotide microarray analysis, or ROMA). They independently reported about 250 and 80 loci as changes in copy number from 39 and 20 normal individuals, respectively. Fig. 2 illustrates the general concept of CNV detection based on two-dye array-CGH. Although the average numbers of CNVs per individual genome were similar in two studies (about 12 CNVs per genome), it should be noted that there was little overlap between the results. This discrepancy between studies was possibly due to the use of different platforms and experimental conditions in different populations. However, it is also probable that there are still large numbers of structural variants that have yet to be discovered (Buckley *et al.*, 2005; Eichler, 2006).

One following study that provided evidence on the widespread presence of large-scale structural variations in the human genome was based solely on in silico analysis (Tuzun *et al.*, 2005). The sequence-level comparison of two independent genome sequences, *i.e.*, one derived from a human genome reference assembly

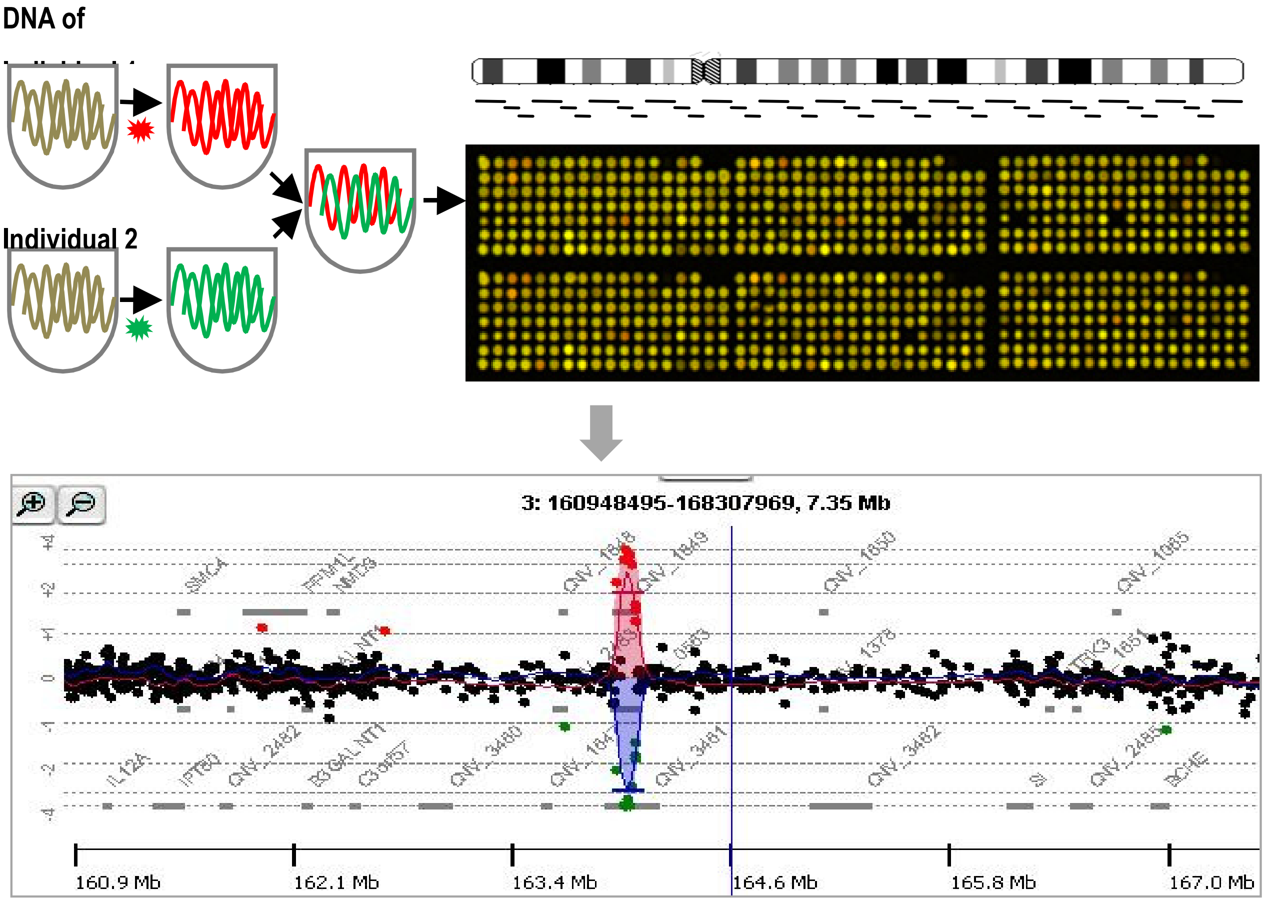

Fig. 2. CNV screening procedure between two individuals using two-dye array-CGH platform. Dye swap hybridization is useful to get reliable interpretation of CNVs. Both individual DNAs are labeled with Cy3 and Cy5, respectively, and co-hybridized. Then, the same experiment is repeated by switching the dyes ('dye-swap'). When both dye-swap profiles show peaks beyond the cutoff and inverted peaks of similar intensity at the same location simultaneously, we interpret them as true CNVs.

and the other from fosmid clones of a genomic library, revealed about 300 structural variations, including inversions. This method can detect various types of structural variants, including inversion, which is not detectable by conventional array-CGH platforms. Indeed, the results by Tuzun *et al.* (2005) can be used as validated control for primary verification or for parameter tuning for the development of CNV-detection platforms or algorithms. Although the use of this method is currently limited by the unavailability of sequence data, ongoing efforts to sequence the individual human genome and to develop cost-effective sequencing platforms (Bennett *et al.*, 2005) will be able to facilitate sequence-level genome comparisons and the identification of highly qualified structural variants in the near future.

Two studies by McCarroll *et al.* and Conrad *et al.*, which focused on the identification of deletion variants (McCarroll *et al.*, 2006; Conrad *et al.*, 2006), used 1.2 million SNP genotyping data from The International HapMap Consortium (International HapMap Consortium. 2005). They assumed that allelic deletion causes the discard of probes in SNP genotyping. For example, the runs of consecutive probes with null genotype calls or runs of SNP genotypes whose allelic frequencies deviate from expected Hardy-Weinberg equilibrium ratios or expected Mendelian inheritance patterns might represent the presence of deleted loci. They independently reported about 600 potential deletions as small as less than 100 bp. The relatively small size of the identified variants, compared with the array-CGH method, is due to the high resolution of the platforms. The use of an SNP-centric array platform can be used to identify linkage disequilibrium (LD) of structural variants with nearby SNPs in a given population. But, the discrepancy in deletions that were identified in the two studies was also noted in spite of using similar HapMap populations and identification methods (Eichler 2006).

Recently, a comprehensive CNV analysis was reported based on high-resolution array platforms, Whole Genome TilePath (WGTP), which used 26,000 large insert clones, and Affymetric GeneChip Human Mapping 500K early access, which used 500,000 SNP oligonucleotides. They identified about 1500 genomic segments as copy number variations or CNVRs (copy number variable regions) consisting of overlapping CNVs from 269 HapMap individuals (Redon *et al.*, 2006). The results from the two platforms are worth comparing becasuse they provide the highest currently achievable resolution and are often selected as primary platforms in many other studies. Firstly, the CNVs that are identified from BAC-based array-CGH are generally larger than those from oligonucleotide-based arrays (230 kb and 80 kb of median size, respectively). This overestimation of CNVs by BAC-based array-CGH is due to the large insert clones that are used, which has been frequently reported (Iafrate *et al.*, 2006). Secondly, the actual boundaries of structural variants can not be determined through BAC-based array-CGH. On the other hand, a more accurate determination of variant boundaries can be achieved through SNP-centric oligonucleotide-based arrays that have an extensive number of oligonucleotides. The SNP-centric platform has additional advantages of accompanying SNP genotype information as a potential variant source, combined with large structural variants and its ability to detect the presence of loss of heterozygosity (LOH) or segmental uniparental disomy (Bruce *et al.*, 2005; Mei *et al.*, 2000). But, the SNP-centric platform also has its disadvantages. In spite of the advanced resolution, the relatively low signal-to-noise ratio of oligonucleotide-based hybridization intensity, compared with large insert clone array, might result in higher false-positive rates. Because most CNVs are subtle changes, this makes the results prone to misclassification of signal intensities and, consequently, to statistical errors.

Sometimes, it is pointed out that the SNP-centric array was originally designed for allelic discrimination and is not appropriate for CNV detection because of biased genomic distribution and sequence composition of spotted probes (McCarroll and Altshuler 2007d). Recently proposed oligonucleotide-based array platforms have been designed for CNV detection specifically without sacrificing the advantage of high resolution, which can be a promising solution for CNV detection in the near future (Barrett *et al.*, 2004).

In identifying CNVs in normal populations, one of the fundamental problems is the lack of a reference genome from which diploid states of sample DNA can be inferred. Unlike the array-CGH-based tumor study in which the normal DNA of the same individual can be used as a reference genome, no single DNA source can present the standardized and universal genome in variant analysis. Often, the pooled genome of several individuals has been used to represent the average genome, while the heterogeneity of the used population might affect the copy number inference step, as shown for examples of X chromosomes. Redon *et al.* and Komura *et al.* adopted the pairwise comparison for ac-

Table 1. Copy number variations and associated diseases/phenotypes

Gene	Location	Susceptible type	Phenotype	Reference
FCGR3	1q23.3	Low copy	Glomerulonephritis	Aitman *et al.*, 2006
FCGR3B	1q23.3	Low copy	Systemic lupus erythematosus, microscopic polyangiitis	Fanciulli M *et al.*, 2007
C4	6p21.32	Low copy	Systemic lupus erythematosus	Yang *et al.*, 2007
CCL3L1	17q12	Low copy	HIV/AIDS susceptibility	Gonzalez *et al.*, 2005
CCL3L1	17q12	High copy	Rheumatoid arthritis	McKinney *et al.*, 2008
UGT2B17	4p13	Low copy	Body testosterone level, prostate cancer risk	Jakobsson *et al.*, 2006
CNTNAP2	7q36.1	Low copy	Schizophrenia and epilepsy	Friedman *et al.*, 2008
AMY1	1p21.1	Variable copy	Starch diet	Perry *et al.*, 2007
DEFB4	8p23.1	Variable copy	Crohn disease	Fellermann *et al.*, 2006
DEFENSIN	8p23.1	Low copy	Autism spectrum disorder	Cho *et al.*, 2008
SHANK3	22q13.3	Low copy	Autism spectrum disorder	Moessner *et al.*, 2007
NRXN1	2p16.3	Low copy	Schizophrenia	Kirov *et al.*, 2008
APBA2	15q13.1	Duplication	Schizophrenia	Kirov *et al.*, 2008

curate inference of copy number states in individual loci, which is noteworthy (Redon *et al*., 2006; Komura *et al*., 2006). In pairwise comparison, the hybridization intensities of one sample is compared with those of all other remaining samples as one large reference, and the diploid states of loci can be more accurately inferred from the multiple comparison results.

Clinical implications of CNVs and disease association study

In spite of recent technological developments of genetic polymorphism-oriented disease association studies, still little is known about the effects of genetic polymorphisms on common complex diseases. One of the ultimate goals in exploring CNVs is to systematically assess the association between such variants and the disease. Although it is unlikely that all CNVs in the human genome are associated with diseases, evidence of the association of CNVs and a wide spectrum of human diseases has rapidly accumulated. Table 1 summarizes the CNVs that have been reported to be associated with diseases. CNVs can affect disease susceptibility or individual differences in responses to drugs through alteration of gene expression. Stranger *et al*.'s and Heidenblad *et al*'s reports coherently showed positive correlations between DNA copy number dosage and gene expression level (Stranger *et al*., 2007; Heidenblad *et al*., 2005). If a CNV region contains transcriptional regulatory elements rather than protein coding genes, it still can affect gene expression levels by changing transcriptional regulation or heterochromatin spread (Reymond *et al*., 2007).

Conclusion

The genomic fraction that is occupied by CNVs is now estimated to be about 600 Mb, already exceeding that of single base-level variants. It is likely that the number of CNVs and the genomic fraction that is affected by structural variants will continue to expand, and many of them will be used for more practical purposes, including disease association or population studies. However, it should be remembered that the current CNV entries are plagued by substantial amounts of false-positive and false-negative results. Only a small portion of them have been validated by independent methods. To overcome this, it is necessary to improve scanning platforms, including optimizing experimental conditions and developing more reliable CNV calling algorithms. In the meantime, it is required for individual researchers to know the characteristics of the available platforms and analytical techniques to use them or to interpret the published results properly.

Acknowledgements

This work is supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0405-BC02-0604-0004) and (01-PJ3-PG6-01GN09-003).

References

Aitman, T.J., R.Dong, T.J., Vyse, P.J., *et al*. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* 439, 851-855.

Barrett, M.T., Scheffer, A., Ben-Dor, A., *et al*. (2004). Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. U. S. A* 101, 17765-17770.

Bennett, S.T., Barnes, C., Cox, A., Davies, L., and Brown, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics* 6, 373-382.

Bruce, S., Leinonen, R., Lindgren, C.M., Kivinen, K., Dahlman-Wright, K., Lipsanen-Nyman, M., Hannula-Jouppi, K., and Kere, J. (2005). Global analysis of uniparental disomy using high density genotyping arrays. *J. Med. Genet.* 42, 847-851.

Buckley, P.G., Mantripragada, K.K., Piotrowski, A., az de, S.T., and Dumanski, J.P. (2005). Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet.* 21, 315-317.

Check, E. (2005). Human genome: patchwork people. *Nature* 437, 1084-1086.

Cheng, Z., Ventura, M., She, X., *et al*. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437, 88-93.

Cho, C.S., Yim, S.H., Yoo, H.K., *et al*. (2008). Copy number variations associated with idiopathic autism identified by whole-genome array-CGH. *Psychiatric Genetics* in press.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* 38, 75-81.

Cooper, G.M., Nickerson, D.A., and Eichler, E.E. (2007). Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* 39, S22-S29.

de Smith, A.J., Tsalenko, A., Sampas, N., *et al*. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* 16, 2783-2794.

Eichler, E.E. (2006). Widening the spectrum of human genetic variation. *Nat. Genet.* 38, 9-11.

Fanciulli, M., Norsworthy, P.J., Petretto, E., *et al*. (2007). FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* 39, 721-723.

Fellermann, K., Stange, D.E., Schaeffeler, E., *et al*. (2006). A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* 79, 439-448.

Feuk, L., MacDonald, J.R., Tang, T., Carson, A.R., Li, M., Rao, G., Khaja, R., and Scherer, S.W. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS. Genet.* 1, e56.

Freeman, J.L., Perry, G.H., Feuk, L., *et al.* (2006). Copy number variation: new insights in genome diversity. *Genome Res.* 16, 949-961.

Friedman, J.I., Vrijenhoek, T., Markx, S., *et al.* (2008). CNTNAP2 gene dosage variation is associated with schizophrenia and epilepsy. *Mol. Psychiatry.* 13, 26126-26126.

Gonzalez, E., Kulkarni, H., Bolivar, H., *et al.* (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 307, 1434-1440.

Heidenblad, M., Lindgren, D., Veltman, J.A., Jonson, T., Mahlamäki, E.H., Gorunova, L., van Kessel, A.G., Schoenmakers, E.F., and Höglund, M. (2005). Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* 24, 1794-1801.

Hinds, D.A., Kloek, A.P., Jen, M., Chen, X., and Frazer, K.A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* 38, 82-85.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949-951.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299-1320.

Istrail, S., Sutton, G.G., Florea, L., *et al.* (2004). Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. U. S. A* 101, 1916-1921.

Jakobsson, J., Ekstrom. L., Inotsume, N., Garle, M., Lorentzon, M., Ohlsson, C., Roh, H.K., Carlström, K., and Rane, A. (2006). Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. *J. Clin. Endocrinol. Metab.* 91, 687-693.

Kirov, G., Gumus, D., Chen, W., *et al.* (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum. Mol. Genet.* 17, 458-465.

Komura, D., Shen, F., Ishikawa, S., *et al.* (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* 16, 1575-1584.

Lupski, J.R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14, 417-422.

McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37-S42.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., *et al.* (2006). Common deletion polymorphisms in the human genome. *Nat. Genet.* 38, 86-92.

McKinney, C., Merriman, M.E., Chapman, P.T., *et al.* (2008). Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann. Rheum. Dis.* 67, 409-413.

Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M.S., Reid, B.J., and Lockhart, D.J. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* 10, 1126-1137.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 16, 1182-1190.

Moessner, R., Marshall, C.R., Sutcliffe, J.S., *et al.* (2007). Contribution of SHANK3 mutations to autism spectrum disorder. *Am. J. Hum. Genet.* 81, 1289-1297.

Perry, G.H., Dominy, N.J., Claw, K.G., *et al.* (2007). Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* 39, 1256-1260.

Przeworski, M., Hudson, R.R., and Di, R.A. (2000). Adjusting the focus on human variation. *Trends Genet.* 16, 296-302.

Redon, R., Ishikawa, S., Fitch, K.R., *et al.* (2006). Global variation in copy number in the human genome. *Nature* 444, 444-454.

Reymond, A., Henrichsen, C.N., Harewood, L., and Merla, G. (2007). Side effects of genome structural changes. *Curr. Opin. Genet. Dev.* 17, 381-386.

Sachidanandam, R., Weissman, D., Schmidt, S.C., *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-933.

Sebat, J., Lakshmi, B., Troge, J., *et al.* (2004). Large-scale copy number polymorphism in the human genome. *Science* 305, 525-528.

Sharp, A.J., Locke, D.P., McGrath, S.D., *et al.* (2005). Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* 77, 78-88.

Stankiewicz, P., and Lupski, J.R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 18, 74-82.

Stranger, B.E., Forrest, M.S., Dunning, M., *et al.* (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315, 848-853.

Tuzun, E., Sharp, A.J., Bailey, J.A., *et al.* (2005). Fine-scale structural variation of the human genome. *Nat. Genet.* 37, 727-732.

Warburton, D. (1991). De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am. J. Hum. Genet.* 49, 995-1013.

Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., *et al.* (2007). A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.* 80, 91-104.

Yang, Y., Chung, E.K., Wu, Y.L., *et al.* (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am. J. Hum. Genet.* 80, 1037-1054.