Cytogenet Genome Res 123:333–342 (2008)
DOI: 10.1159/000184725

Cytogenetic and Genome Research

Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels

J.H. Lee[a] J.T. Jeon[b]

[a]Division of Animal Science and Resources, College of Agriculture and Life Sciences, Chungnam National University, Daejeon and [b]Division of Applied Life Science, Gyeongsang National University, Jinju (Korea)

Accepted in revised form for publication by H. Kehrer-Sawatzki and D.N. Cooper, 18 June 2008.

Abstract. Copy number variations (CNVs) have effects on phenotypes by altering transcription levels of genes and may have major impacts on protein sequence, structure and function. Therefore, CNV screening and analysis focused on the identification of CNV-genetic disease relations are actively progressing. CNVs can be detected and analyzed by various methodologies at the genome-wide and locus-specific levels. The genome-wide analysis of CNVs has been enhanced by bioinformatic tools for long-range sequence analysis, and comparative genome hybridization using microarrays containing either single nucleotide polymorphisms or bacterial artificial chromosome clones that represent the whole genome. RFLP followed by Southern blot analysis, quantitative real-time PCR, pyrosequencing, ligation detection reaction and the invader assay have become the main tools for locus-specific analysis so far. In this review, we present a brief principle, application history, and strengths and weaknesses of the methods used to detect CNVs at the genome-wide and locus-specific levels.

There is a wide range of genetic variations, from single nucleotide polymorphisms (SNPs) to chromosomal abnormalities that include the well-known example of Down syndrome in humans (Jacobs et al., 1959). Phenotypic variation due to the copy number of the *Bar* gene in *Drosophila melanogaster* was discovered in 1936 (Bridges, 1936). Ever since the human genome sequencing results were released to the public domain, many reports have described the copy number variations (CNVs) of DNA segments, with sizes ranging from kilobases (kb) to megabases (Mb). These CNVs, which contain deletions, insertions and multi-site variants, have been discovered in mammals, such as human, mouse, rat, rhesus macaque and chimpanzees (Perry et al., 2006; Redon et al., 2006; Graubert et al., 2007; Guryev et al., 2008; Lee et al., 2008). The species mentioned above already have draft genome sequences, and the genome-wide CNVs were investigated. This indicates that all mammals have CNVs and they are important for genetic variability and evolution.

Current CNV research, mostly in humans, is focused on the identification of relationships between CNVs and genetic diseases. The importance of CNVs in human genetic disorders became evident in reports that over 300 proven disease-causing genes overlapped with CNVs (Sebat et al., 2004) and that 14.5% of CNVs overlapped with the OMIM Morbid Map (Redon et al., 2006). This research on CNVs will eventually be extended to various areas in order to delineate complex genetic phenomena. CNVs can affect phenotypes by altering transcription levels of genes (Hegele, 2007). In addition, they may have major impacts on protein sequence, structure and function because expansions of gene family size by gene duplication occurs in several mammalian species (Demuth et al., 2006), and there is a significant increase in the effective SNP rate in duplication-transition regions when compared with unique genomic sequences (She et al., 2006).

These CNVs can be detected and analyzed by methodologies targeting the whole genome (referred to as genome-wide level in this review), as well as those restricted to a certain location on the chromosomes (locus-specific level).

This work was supported by grants from the BioGreen 21 program (20070401-034-029-01 and 20050301-034-467), Rural Development Administration, Republic of Korea.

Request reprints from Jin-Tae Jeon
Division of Applied Life Science, Gyeongsang National University
900 Gajwadong, Jinju, Gyeongnam 660-701 (Korea)
telephone: +82 55 751 5516; fax: +82 55 756 7171
e-mail: jtjeon@gnu.ac.kr

1424–8581/08/1234–0333$24.50/0

Accessible online at:
www.karger.com/cgr

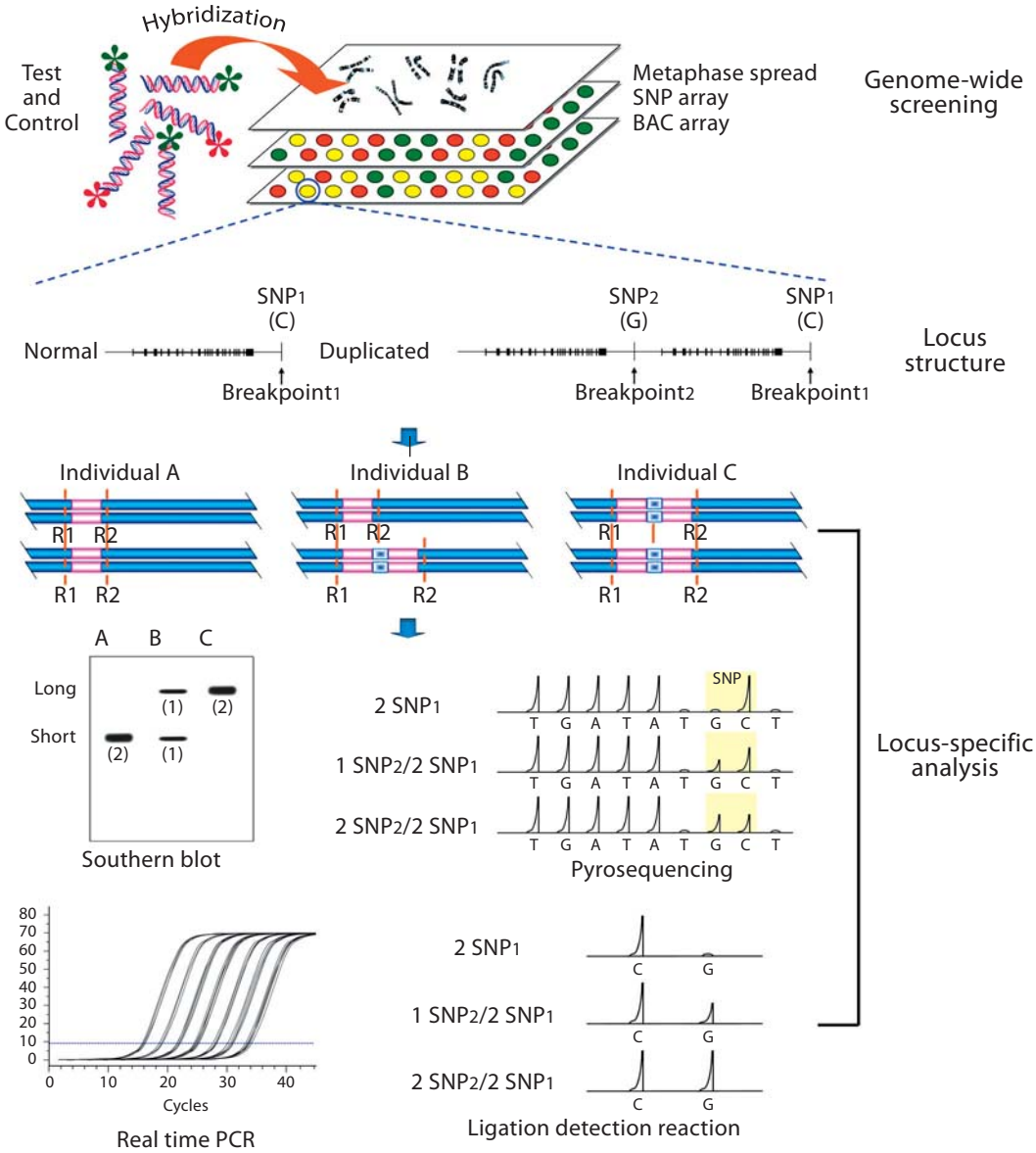

Fig. 1. An illustration of the flow from genome-wide screening to locus-specific analyses of CNVs. Genome-wide analysis using hybridization of test and control samples on microarrays containing either SNPs or large genomic clones. Once a CNV of interest is identified at the genome level, the genomic organization and internal sequence information must be elucidated. Breakpoint 1 and 2 indicate duplication initiation and stop point respectively. Nucleotides G and C indicate the representative SNPs for Breakpoint 1 and 2 respectively. The methodology for locus-specific analysis could be planned on the basis of the information. A CNV needs to be analyzed more precisely at the locus level, using a variety of techniques. RFLP followed by Southern blot can be applied for a long-range mapping (from R1 to R2). Quantitative real-time PCR, oligonucleotide ligation assay and pyrosequencing can analyze copy numbers using the representative SNPs.

Genome-wide analysis of CNVs has been enhanced by a comparative genome analysis using bioinformatics tools with long-range sequences (She et al., 2006). Hybridization of test and control samples on microarrays containing either SNPs or large genomic clones has revealed that repetitive elements and motifs with unknown function, as well as functional segments of genes, are spread throughout the genome. As illustrated in Fig. 1, once a CNV of interest is identified at the genome level, it needs to be analyzed more precisely at the locus level, and ultimately, the genotype and haplotype must be determined to elucidate its relationship with a particular genetic alteration (Seo et al., 2007). Locus-specific CNVs were identified in conjunction with genome-wide screening (Iafrate et al., 2004; Sharp et al., 2005; Wong et al., 2007) and independently through gene family studies (Ghanem et al., 1988; Trask et al., 1998) or functional analysis of genes associated with a certain phenotype (Johansson Moller et al., 1996). Among the latter cases, genomic organization and detection methods for *C4* CNVs associated with systemic lupus erythematosus in the HLA class III

region and for *KIT* CNVs related to porcine Dominant White phenotype (*Dominant White/KIT* locus) are well established; therefore, these regions could be suitable model loci for explaining and comparing methodologies used to detect locus-specific CNVs in general. The human *C4* CNV belongs to a tandem duplication having bimodular form (spanning about 120 kb) of the *RP–C4-CYP21-TNX* (RCCX) region on human chromosome (HSA) 6p21.3. Long-range mapping using rare-cut enzyme digestion and electrophoresis of high molecular weight DNA followed by Southern hybridization (RFLP-Southern blot) has been used for duplication analysis of the long module (Yang et al., 2007). Also, quantitative real-time PCR (qPCR) has been used for a detailed survey of the CNVs of *C4B* and *C4A*, *C4L* and *C4S*, and the RCCX module (Wu et al., 2007). The porcine *KIT* CNV is also a tandem duplication and is located on pig chromosome (SSC) 8p11. To analyze the *KIT* locus, RFLP-Southern blot (Marklund et al., 1998), minisequencing, qPCR (Pielberg et al., 2002), invader assay, pyrosequencing (Pielberg et al., 2003) and quantitative oligonucleotide ligation assay (qOLA, Seo et al., 2007) have been used. As mentioned briefly, a variety of techniques is necessary to provide in-depth analysis of total copy numbers, SNPs on paralogs, breakpoints, etc. on the locus level.

In this article, we review the development and current state of the methodology for analysis of CNVs and discuss perspectives of these techniques.

Methodology used for genome-wide detection of CNVs

The genome-wide CNVs can largely be detected by two DNA chip-based methods that were comprehensively compared by Redon et al. (2006). We call these (1) CGH-based CNV detection and (2) SNP array-based CNV detection. The differences between the two methods lie in the type of microarray used and also the hybridization methods. The former uses two fluorescent dyes for labeling test and reference samples, and the samples can be hybridized to the same microarray spot. On the other hand, the latter method uses only one fluorescent dye for each sample, and comparison between the samples can give the location of the CNV. Basically, both methods use $\log_2$ ratios for identification of the locations of CNVs. When different CNV detection techniques were applied in the same samples, it requires the development of distinct algorithms to identify CNVs.

Comparative genomic hybridization(CGH)-based CNV detection

The CGH method is essentially based on the fluorescence in situ hybridization (FISH) method, and the probes can be hybridized to the complementary DNA sequence in metaphase spreads of chromosomes. In comparison with FISH, which is normally used as a mapping tool for identifying the cytogenetic location of a gene or a DNA segment using fluorescence dyes, the CGH method uses two different fluorescent dyes for the test (unknown or experimental) and reference DNA samples, respectively. By measuring the fluorescence signals for the two dyes and assuming that the reference sample has a normal, diploid copy number, the CNVs can be detected as either the gain or loss of signal in the test sample. This CGH method was originally developed for detecting copy number variations and their chromosomal locations in tumor and normal samples (Kallioniemi et al., 1995). The major limitation of this technique is the relatively low resolution, normally more than 5 Mb of detectable signal. In order to overcome this size limitation, array-based CGH has been developed, benefiting from the improvements in microarray technology over the last decade. The sources of the DNA used on the microarray were BAC clones, cDNA clones, oligonucleotides and genomic PCR products (Carter, 2007). The large-insert clones from human libraries, such as bacterial artificial chromosome (BAC) clones, cover most of the human genome, and this method has been called array CGH (Pinkel et al., 1998; Cheung et al., 1999). Later on, genome-wide tiling-path human BAC arrays became available (Ishkanian et al., 2004; Fiegler et al., 2006). The advantage of using the large insert clones for array CGH is not only the improvement of resolution, but also the ease of extracting the sequences of the CNV regions and drawing conclusions about the phenotypic variations. Recently, array CGH has been widely used for identifying CNVs and has become a standard method for detecting tumors (Carter, 2007). Until now, CNV research has been focused on the identification of relationships with genetic diseases. However, these CNVs are not always linked with genetic diseases, based on the observations of Iafrate et al. (2004) and Sebat et al. (2004) that undiseased individuals have these CNVs, indicating the complexity of mammalian genomes including the human genome.

cDNA microarrays are mostly used for characterizing variation in gene expression, identifying tissue-specific expression of genes and identifying up- or down-regulated genes in a certain environment (Schena et al., 1995). These cDNA microarrays have also been used to detect genome-wide CNVs, and criteria for determining CNVs have been developed (Pollack et al., 1999; Park et al., 2006). However, cDNA microarray-based CNV detection has two limitations. One is that genes are unevenly distributed in the genome, and therefore this method cannot cover the whole genome, meaning that some of the CNVs may be missed in the analysis. The other limitation is due to hybridization problems. The cDNA clones on the microarray do not have introns, so any mismatches between genomic DNA and cDNA will affect the hybridization signals and ultimately changes in relative ratios, which can give more possibility of errors in the final results.

As a result of the development of microarray technology, along with powerful bioinformatics tools, oligonucleotide arrays are now commercially available for CNV detection. Two different technologies were used for making the oligonucleotide microarray slides. Inkjet technology has been used for making a 44K human genome CGH microarray by Agilent Technologies, based on 60-mer oligonucleotide probes spaced at an average of 35 kb (http://www.home.ag-

Table 1. Example of locus-specific CNVs and the methods applied for the detection

Locus	Chromosome location	Related function/disorder	Applied method	Reference
DEFA/*DEFB*	HSA8p23–p22	Antimicrobial mediator of innate immunity	qPCR	Linzmeier and Ganz (2005), Chen et al. (2006)
			RFLP - Southern blot	Aldred et al. (2005)
FCGR2, FCGR3	HSA1q23	Glomerulonephritis and autoimmunity	Multiplex ligation-dependent probe amplification (MLPA)	Breunis et al. (2008)
			qPCR	Aitman et al. (2006), Fanciulli et al. (2007)
CYP2D6	HSA22q13	Debrisoquine metabolism	Pyrosequencing	Zackrisson and Lindblom (2003)
			Invader assay	Nevilie et al. (2002)
			qPCR	Müller et al. (2003), Schaeffeler et al. (2003)
			PCR - RFLP	Gjerde et al. (2008)
RCCX	HSA6q21	Systematic lupus erythematosus	RFLP - Southern blot	Yang et al. (2007)
			Locus-specific PCR	Lee et al. (2006)
			qPCR	Wu et al. (2007)
GSK3B	HSA3q13	Bipolar disorder	qPCR	Lachman et al. (2007)
CCL3L1	HSA17q21	Susceptibility to HIV-1 infection	qPCR	Nakajima et al. (2007)
GSTM1	HSA1p13	Atopic asthma	qPCR	
GSTT1	HSA22q11			Brasch-Andersen et al. (2004)
KIT	SSC8p11	Coat color (Dominant white)	RFLP – Southern blot	Johansson Moller et al. (1996)
			qPCR	Johansson Moller et al. (1996), Pielberg et al. (2002)
			Minisequencing	Pielberg et al. (2002)
			Invader assay, Pyrosequencing	Pielberg et al (2003)
			Quantitiative oligonucleotide ligation assay (qOLA), Pyrosequencing	Seo et al. (2007)

ilent.com/). This microarray was used to detect hidden aberrations in patients with myelodysplastic syndromes (Evers et al., 2007). Currently, a 244K human genome CGH microarray is commercially available through Agilent Technologies and has been used in various research areas. Another technology for making oligonucleotide microarrays is photolithography. NimbleGen Company commercialized a 385K human tiling array and a 2.1-million oligonucleotide array (HD2) using this technology, which has been used for various lines of CNV research (http://www.nimblegen.com/). The main limitation of these oligonucleotide arrays is the poor signal-to-noise ratio of hybridizations, and researchers normally use the average from only a few separate hybridizations in order to decrease the measurement variance (Carter, 2007).

CNV detection using SNP arrays

A large portion of phenotypic variation is due to nucleotide changes, and the investigation of SNPs has become a very important research area. In the case of humans, more than 1 million SNPs were reported from four of the major human populations (International HapMap Consortium, 2005), and recently, a second generation of the human haplotype map containing over 3.1 million SNPs was reported (International HapMap Consortium, 2007). Using these large numbers of SNPs, two companies, Affymetrix and Illumina, have commercialized high-throughput human SNP arrays, mainly used for detecting the causative genes for specific diseases and traits. For the Affymetrix SNP array, each SNP allele is determined by the results of 20 matched and mismatched probe pairs (http://www.affymetrix.com/). On the other hand, Illumina uses a different platform called a bead array (http://www.illumina.com/). These two different SNP arrays have both been used for CNV detection. In these SNP arrays, only one test DNA sample is labeled with fluorescent dye, which is different from the array CGH method. In order to identify the CNVs, the SNP information is compared in two different individuals. The limitation of these arrays is that the SNPs are not evenly distributed across the genome, and additionally the CNV regions are difficult to genotype.

Technologies used to analyze locus-specific CNVs

In Table 1, examples of locus-specific CNVs and detection methods are shown. Reported locus-specific CNVs can be classified into two main types: 1) module duplications in which a few different kinds of genes/segments constitute a module; and 2) unit duplications composed of paralogs originating from one type of ancestral gene/segment. Because the *C4* region CNV in HLA class III (module CNV)

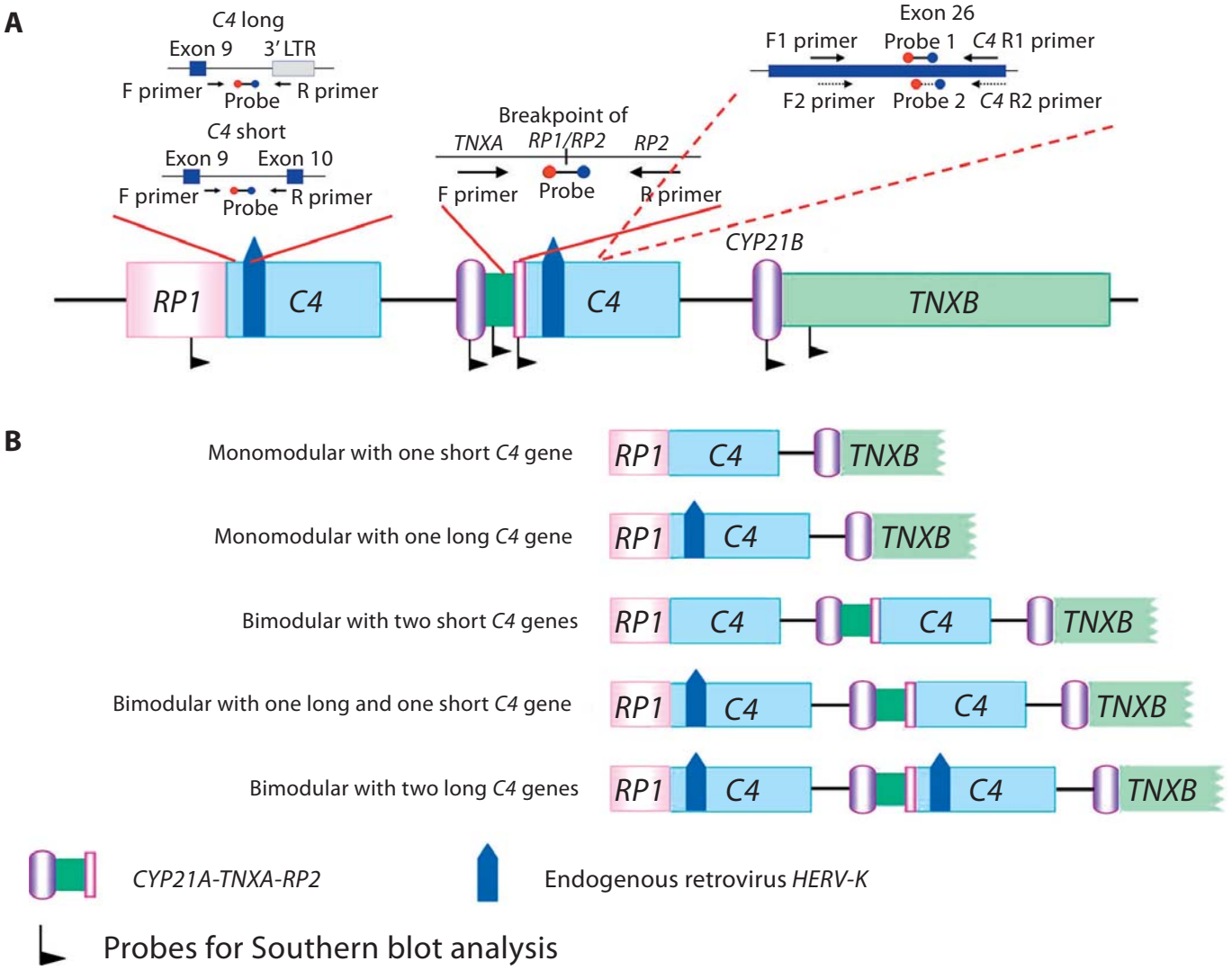

Fig. 2. A schematic description of the RCCX modules and haplotypes in HLA class III on HSA6q21. (**A**) The map showing the gene organization with a bimodular haplotype containing two long *C4* genes. The target positions of quantitative real-time PCRs are indicated over the map. Flags turning upside down below *RP1*, *RP2*, *CYP21A*, *CYP21B*, *TNXA* and *TNXB* indicate the location of DNA probes employed for RFLP followed by Southern blot analysis. (**B**) Schematic descriptions of RCCX haplotypes. Each halpotype could be estimated by a long-range mapping for the total length of the module, and constituent analysis using gene/segment specific probes or qPCR for the total *C4* copies.

and the *KIT* CNV in pigs (unit CNV) are well characterized, and applied methods are diverse, we will primarily explain and discuss the techniques used for these two model loci.

In Fig. 2A, the RCCX module, which is composed of *RP*, *C4*, *CYP21*, *TNX*, and gene segments, is illustrated. Derived haplotypes become more complicated by internal involvement of endogenous retroviral HERV-K sequences in the *C4* gene, generating long (*C4L*) and short (*C4S*) forms (Fig. 2B). More detailed information about SNPs on *C4A* and *C4B*, their functions and CNV-disease relations is described in another paper in this special issue by Szilágyi and Füst.

Figure 3A shows a schematic description of tandem duplication at the porcine *KIT* locus. Two *KIT* mutations cause the Dominant White phenotype in pigs: 1) gene copy numbers associated with a semi-dominant phenotype, which is designated as normal and duplicated (Fig. 3A), and 2) a splice mutation leading to the fully dominant allele, which is depicted as an SNP (G/A) at the first nucleotide of intron 17. In Fig. 3B, four known major alleles at the *KIT* locus are described: the recessive *i* allele for the Color phenotype, the I^P allele for the Patch phenotype, the dominant *I* allele for the White phenotype and I^{Be} for the Belt phenotype. Diversity of the dominant *I* allele has been reported as I^1, I^2, I^3 and I^L by Pielberg et al. (2003). All possible alleles and their genomic organizations are shown in Fig. 3B.

RFLP followed by Southern blot analysis (RFLP – Southern blot)

The most conventional method for interrogating CNVs in the range of 5–500 kb is RFLP – Southern blot. Isolation

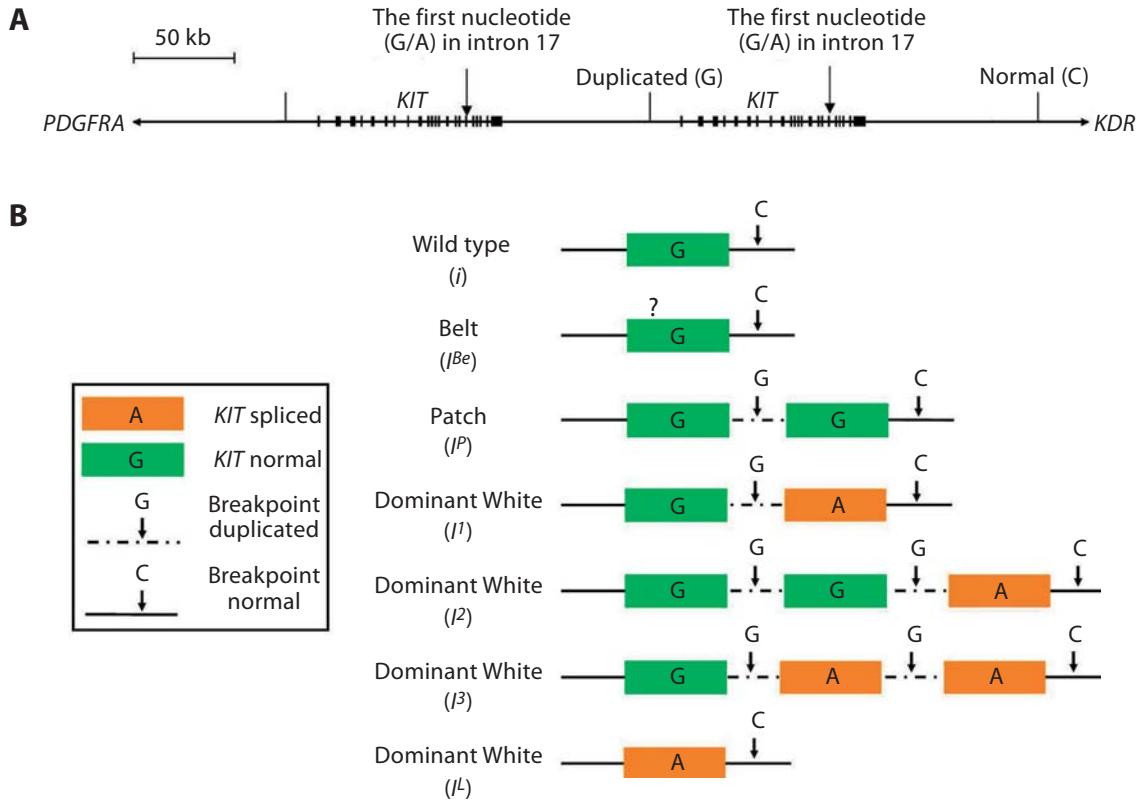

Fig. 3. A schematic description of tandem duplication at the *KIT* locus on SSC8p11. (**A**) The duplication unit is about 450 kb. A breakpoint at the junction between the two *KIT* copies is designated as duplicated and nucleotide G, and another breakpoint at the distal end of the 2nd *KIT* copy is designated as normal and nucleotide C. A splice donor mutation leading to the fully dominant allele, which is marked as an SNP(G/A) at the first nucleotide of intron 17 is indicated by an arrow. (**B**) Schematic descriptions of *KIT* alleles. A question mark in the I^{BE} allele means an unidentified polymorphism causing the Belt phenotype. Discrimination between *i* and I^{BE} is not possible at present.

of high molecular weight DNA, digestion with rare-cut enzyme, and resolving digested DNA fragments by pulsed field gel or low percentage (< 1%) agarose gel electrophoresis are sequentially applied for this method. The resolved fragments are transferred to membranes and hybridized with appropriate probes.

RFLP analysis of the *RCCX* locus was achieved by long-range mapping using either *Pme*I digestion to catch the total length of the locus (*BF* to *TNXB*) or *Taq*I digestion to address what kinds of components are located within the duplicated module (Yang et al., 2007). A *C4*-specific probe was used for long-range sizing. Probes specific to *RP1*, *RP2*, *CYP21A*, *CYP21B*, *TNXA* and *TNXB* were used for the component analysis through Southern hybridization. Integrating the results from the two analyses, RCCX modules and gene CNVs were clearly resolved in four selected families with systemic lupus erythematosus (Yang et al., 2007). However, the analysis did not use direct quantification of the CNV but instead detected the CNV state through combining and finding pieces that fit every component in the module. Johansson Moller et al. (1996) applied the quantitative RFLP-Southern blot method to analyze the porcine *KIT* CNV. *Taq*I-digested DNAs were resolved on a 0.7% agarose gel, transferred onto membranes and hybridized with probes prepared from *KIT* intron 18 and *PLANH2*. The probes were separated, labeled and applied together in the hybridization. The signals from both probes were quantified using an image analyzer. Detection of CNV was achieved by calculating the ratio between the signals from both probes. Discrimination of CNV from two to four *KIT* copies in a diploid genome was successful.

The RFLP-Southern blot method may be a good choice to resolve large size CNVs and structural variations within the duplication. As this method does not require a high-priced detection apparatus, it could become a widely-used method. However, there are some drawbacks to this method: 1) the higher cost per analysis compared to other methods such as real-time PCR, pyrosequencing, etc.; 2) time-consuming and laborious procedures that require more than a week between DNA isolation and detection; 3) the need for purified high molecular weight DNA; and 4) the potential need for radioisotopes for detection.

Quantitative real-time PCR (qPCR)

qPCR is performed on an apparatus uniting a thermal cycler and an optical instrument to capture spectral fluo-

rescence, and it can monitor the kinetics of the entire PCR in real time. qPCR does not require post-PCR processing of PCR products. It helps high-throughput analysis by reducing the chances of carryover contamination (Dorak, 2006). At present, therefore, it has become one of the most popular and effective methods for analyzing CNVs (Chen et al., 2006; Fanciulli et al., 2007; Nakajima et al., 2007; Wu et al., 2007).

Fluorescence-monitoring systems for qPCR consist of the following: 1) hydrolysis probes; 2) hybridizing probes; and 3) DNA-binding agents. The TaqMan assay is a typical system using the hydrolysis probe, for which fluorescence resonance energy transfer (FRET) technology is applied (Hiyoshi and Hosoi, 1994; Chen et al., 1997). The probe contains two labels, a fluorescent reporter dye at the 5′-end and a fluorescent quencher at the 3′-end. The probe is hybridized to the target sequence prior to annealing PCR primers. When the polymerase elongates the newly synthesized strand, the probe is cleaved by the 5′-exonuclease activity of the polymerase (Holland et al., 1991). When the reporter and quencher are separated, the reporter dye is activated. Hybridizing probes also use the FRET technology, but these differ from hydrolysis probes because they contain two short probes that bind to an internal sequence of the amplified fragment. The donor dye labeled at the 5′-end of one probe is excited by the light source of a qPCR instrument and transfers part of this excitation energy to the acceptor/reporter dye labeled at the 3′-end of another probe. Consequently, when two probes are hybridized together to the target PCR product, the fluorescence signal can be captured. The use of a non-sequence-specific fluorescent intercalating agent in qPCR is cheaper and simpler than the two probe types mentioned above (Dorak, 2006). The agent binds to double-stranded DNA. A fluorogenic minor groove binding dye, SYBR green, exhibits little fluorescence when in solution but emits a strong fluorescent signal upon binding to double-stranded DNA (Morrison et al., 1998). With the accumulation of target sequence during PCR, more probes and agents are hydrolyzed, hybridized or intercalated, and the fluorescence signal increases. Quantification of the initial amount of template in a reaction is based on the number of cycles up to a threshold (C_T) at which the fluorescence signal is exponentially increased and passes an arbitrarily defined value (Livak et al., 1995).

Wu et al. (2007) analyzed *C4* copy numbers and the constituents in the RCCX module by qPCR using TaqMan probes. For *C4* CNV analysis, primers on the *TNXA* and *RP2* segments were used for PCR, and a hydrolysis probe designed by using the junction sequence of the breakpoint between the two segments was applied for the TaqMan assay. They showed that resolving two to four total copies and identifying the copy number of each type of *C4* were both possible. PCR primers and TaqMan probes specific to *C4A*, *C4B*, *C4L* and *C4S* were applied for the constituent analysis. *RPI* was used as an endogenous control for the analysis. qPCR with TaqMan probes was used for *KIT* CNV analysis (Pielberg et al., 2002). The assay was carried out for the target *KIT* gene and a single copy control sequence from *ESR*. They inferred *KIT* copy numbers using the equation C_T(ESR) – C_T(KIT) and estimated the relative copy number from two to five copies.

In general, qPCR using TaqMan probes requires very little optimization. It presents reliable results because nonspecific amplified PCR products and primer dimers do not interfere with the assay (Dorak, 2006). However, we have found a few limitations in this method: 1) underestimation with increasing copy number (a question of accuracy); 2) incrementing of the standard deviation with increasing copy number (a question of precision); 3) applications restricted to human and animal samples having pedigree information so far; and 4) need of good controls to yield reliable standard curves and convert to real copy numbers. Underestimation at higher copy numbers seems to be a common phenomenon that could be overcome through the use of a well-approximated standard curve. Because a good curve should result from the choice of a proper control sequence, selection of a good control would be the most important factor for the qPCR method.

Pyrosequencing

The principle of pyrosequencing was developed by Ronaghi et al. (1998). It is based on the detection of released pyrophosphates (PP_i) upon addition of one of four dNTPs corresponding to the template sequence at the 3′ end of the newly synthesized strand. This technology uses an enzyme cascade of DNA polymerase, sulfurylase, luciferase and apyrase. The PP_i released during DNA synthesis is converted quantitatively to ATP by sulfurylase in the presence of adenosine 5′-phosphosulfate. The ATP is used for the luciferase-mediated conversion of luciferin to oxyluciferin, generating visible light in amounts that are proportional to the amount of ATP. Quantification for CNV analysis simultaneously with sequence analysis for SNP detection is therefore plausible.

Pielberg et al. (2003) applied this method to analyze *KIT* copy numbers and to distinguish between wild and mutant copies. The estimation of the genotype of animals was performed using the combined information from the two pyrosequencing analyses. Two to six *KIT* copies were resolved, and the ratio of mutant copy was identified in the total copy number. Zackrisson and Lindblom (2003) used this method for the analysis of *CYP2D6* CNVs.

There are a few drawbacks to this method: 1) it is not real-time but end-point analysis. PCR followed by this method must be well-optimized and the total number of PCR cycles must be well-established in order to get templates for end-point analysis from an exponential phase before reaching a plateau; 2) there is underestimation with increasing copy numbers, as happens in the qPCR method; and 3) there is a strong negative effect resulting from poor quality of template DNA.

Ligation detection reaction (LDR)

An oligonucleotide ligation assay (OLA) had been established by Landegren et al. (1998). When two oligonucleotide probes are hybridized and perfectly base-paired to the sin-

gle-stranded template DNA such that the 3′-end of one probe is immediately adjacent to the 5′-end of the other one, DNA ligase can covalently link these two oligonucleotides (Wu and Wallace, 1989). If a mismatch at either the 3′-end of the first probe or the 5′-end of the second probe is introduced, ligation does not occur. This feature could be applied to detect SNPs. Involvement of thermostable DNA ligase in the assay allowed for product capture and detection in a chain reaction manner. A ligation chain reaction (LCR) based on an OLA has been developed to both amplify DNA and identify any SNPs. Exponential amplification of ligation is achieved by thermal cycling of ligation reactions in the presence of a second set of adjacent oligonucleotides, complementary to the first set and the target. A single-base mismatch prevents ligation/amplification and is thus detected (Barany, 1991). As a result of the introduction of fluorescence reporter dyes and automated DNA sequencers to LDR, the method is done in a multiplex and high-throughput manner.

Seo et al. (2007) used quantitative OLA (qOLA) for the detection of *KIT* CNVs. The first oligonucleotide probe, the common probe, for qOLA was designed from the –24 to the –1 nucleotide position from the duplication breakpoint (Fig. 3A). One of each breakpoint-specific probe, the second probe, for the duplicated or normal copy was designed from the duplication breakpoint (Fig. 3A). The 5′-end of the first probe was labeled with a fluorescent dye, a phosphate group was added to the 5′ ends of the two second probes, and $(dA)_{10}$ and $(dA)_{15}$ were added at the 3′-ends of the second probes to distinguish the two OLA products and separate them further from the unused oligonucleotide peak at the size fractionation step. A cycled ligation reaction for qOLA can be applied because the Ampligase enzyme (Epicentre Biotech, USA) is a thermostable DNA ligase, and the OLA products are resolved on an ABI Prism 3100 Genetic Analyzer (Applied Biosystems, USA). Seo et al. (2007) showed that the qOLA method is a reliable assay for measuring *KIT* CNV from two to six copies and could be used for a variety of samples, such as those in a known pedigree, those with predictable segregation, those without pedigree information, and those consisting of poor-quality genomic DNA.

This study suggested that the application of LDR to a quantification assay rather than positive/negative screening could yield another robust method for analyzing CNVs. However, it is an end-point analysis and needs well-optimized PCR conditions, as did pyrosequencing.

Other techniques

The invader assay uses an allele-specific primary probe that contains a 5′-flap that is non-complementary to the target DNA and an Invader probe hybridized in tandem to the target DNA to form a specific overlapping structure. The Cleavase enzyme cuts the 5′-flap on the primary probe at the base of the overlap with the Invader probes. The cleaved 5′-flap is released as a target-specific product. The target-specific 5′-flap oligos are involved in a secondary reaction, where they act as Invader probes on a FRET cassette, leading to the formation of an overlapping structure that is recognized by the Cleavase enzyme.

As described in Table 1, human *CYP2D6* and porcine *KIT* CNVs have been analyzed using this technology. This method has a limitation in accurately resolving copy numbers greater than three copies of *CYP2D6* (Nevilie et al., 2002). Pielberg et al. (2003) reported that the quantification of *KIT* CNVs using the Invader technology showed a good correlation with that of pyrosequencing. However, as copy number increased, the results of this method became increasingly underestimated.

Locus-specific PCR and multiplex ligation-dependent probe amplification (MLPA) have been applied for the *RCCX* CNVs and *FRGB* CNVs, respectively.

Perspectives

For the genome-wide CNV detection, arrays using large insert genomic clones and SNPs are widely used. However, there are large limitations for using the SNP arrays, since the SNPs are not evenly distributed in the entire genome. In order to overcome this limitation, a genome-wide Affymetrix array (Genome Wide SNP Array 6.0) is already commercially available, which contains non-polymorphic probes for the locations of known CNV regions as well as the genomic locations that were not well-covered with known SNPs. Very recently, an Nsp CN (Copy Number) array was designed using non-polymorphic oligonucleotide probes for the improved detection of global CNV (Shen et al., 2008). Therefore, CNV researches will be more focused on the previously unidentified chromosomal regions using the non-polymorphic probes.

When the CNV has been detected by the array based method, sequencing of the CNV region has to be carried out in order to confirm the identified CNV. Until now, the capillary-based sequencing is predominantly used with high sequencing cost. Recently, low cost and efficient sequencing technologies, such as 454 Life Sciences and Solexa, have been developed and more results of CNV research will be achieved.

A major problem in the measurement of CNVs at the locus level is that the higher the copy number in a genome, the more difficult it is to resolve the copies. There has been underestimation of the copy numbers in most techniques described in this review when copy number increased. The bias would be mainly caused by differences in the kinetics at the early and late stages of PCR amplification. To overcome this limitation, development of more advanced technologies using highly sensitive reporter molecules and detection tools so as to exclude PCR amplification is needed.

Based on current knowledge and developed technologies, these CNVs are known to be very important for understanding diseases and phenotypic variances. In the future, CNV research will be more actively pursued and will give a more detailed understanding of these variations.

References

Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, et al: Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. Nature 439:851–855 (2006).

Aldred PM, Hollox EJ, Armour JA: Copy number polymorphism and expression level variation of the human alpha-defensin genes *DEFA1* and *DEFA3*. Hum Mol Genet 15:2045–2052 (2005).

Barany F: Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Natl Acad Sci USA 88:189–193 (1991).

Brasch-Andersen C, Christiansen L, Tan Q, Haagerup A, Vestbo J, Kruse TA: Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of *GSTM1* and *GSTT1* gene copy numbers. Hum Mutat 24:208–214 (2004).

Breunis WB, van Mirre E, Bruin M, Geissler J, de Boer M, et al: Copy number variation of the activating *FCGR2C* gene predisposes to idiopathic thrombocytopenic purpura. Blood 111:1029–1038 (2008).

Bridges CB: The Bar 'gene': a duplication. Science 83:210–211 (1936).

Carter NP: Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–21 (2007).

Chen Q, Book M, Fang X, Hoeft A, Stuber F: Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR. J Immunol Methods 308: 231–240 (2006).

Chen X, Zehnbauer B, Gnirke A, Kwok PY: Fluorescence energy transfer detection as a homogeneous DNA diagnostic method. Proc Natl Acad Sci USA 94:10756–10761 (1997).

Cheung VG, Dalrymple HL, Narasimhan S, Watts J, Schuler G, et al: A resource of mapped human bacterial artificial chromosome clones. Genome Res 9:989–993 (1999).

Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW: The evolution of mammalian gene families. PLoS ONE 1:e85 (2006).

Dorak MT: Real-Time PCR (Advanced Methods Series) (Taylor & Francis Ltd, Oxford 2006).

Evers C, Beier M, Poelitz A, Hildebrandt B, Servan K, et al: Molecular definition of chromosome arm 5q deletion end points and detection of hidden aberrations in patients with myelodysplastic syndromes and isolated del(5q) using oligonucleotide array CGH. Genes Chromosomes Cancer 46:1119–1128 (2007).

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, et al: *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721–723 (2007).

Fiegler H, Redon R, Andrews D, Scott C, Andrews R, et al: Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res 16:1566–1574 (2006).

Ghanem N, Uring-Lambert B, Abbal M, Hauptmann G, Lefranc MP, Lefranc G: Polymorphism of MHC class III genes: definition of restriction fragment linkage groups and evidence for frequent deletions and duplications. Hum Genet 79:209–218 (1988).

Gjerde J, Hauglid M, Breilid H, Lundgren S, Varhaug JE, et al: Effects of CYP2D6 and SULT1A1 genotypes including SULT1A1 gene copy number on tamoxifen metabolism. Ann Oncol 19: 56–61 (2008).

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, et al: A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet 3:e3 (2007).

Guryev V, Saar K, Adamovic T, Verheul M, van Heesch SA, et al: Distribution and functional impact of DNA copy number variation in the rat. Nat Genet 40:538–545 (2008).

Hegele RA: Copy-number variations add a new layer of complexity in the human genome. CMAJ 176:441–442 (2007).

Hiyoshi M, Hosoi S: Assay of DNA denaturation by polymerase chain reaction-driven fluorescent label incorporation and fluorescence resonance energy transfer. Anal Biochem 221:306–311 (1994).

Holland PM, Abramson RD, Watson R, Gelfand DH: Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Natl Acad Sci USA 88:7276–7280 (1991).

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al: Detection of large-scale variation in the human genome. Nat Genet 36:949–951 (2004).

International HapMap Consortium: A haplotype map of the human genome. Nature 437:1299–1320 (2005).

International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861 (2007).

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, et al: A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36:299–303 (2004).

Jacobs PA, Baikie AG, Court Brown WM, Strong JA: The somatic chromosomes in mongolism. Lancet 1:710 (1959).

Johansson Moller M, Chaudhary R, Hellmén E, Höyheim B, Chowdhary B, Andersson L: Pigs with the dominant white coat color phenotype carry a duplication of the *KIT* gene encoding the mast/stem cell growth factor receptor. Mamm Genome 7:822–830 (1996)

Kallioniemi A, Kallioniemi OP, Citro G, Sauter G, DeVries S, et al: Identification of gains and losses of DNA sequences in primary bladder cancer by comparative genomic hybridization. Genes Chromosomes Cancer 12:213–219 (1995).

Lachman HM, Pedrosa E, Petruolo OA, Cockerham M, Papolos A, et al: Increase in *GSK3* beta gene copy number variation in bipolar disorder. Am J Med Genet B Neuropsychiatr Genet 144:259–265 (2007).

Landegren U, Kaiser R, Sanders J, Hood L: A ligase-mediated gene detection technique. Science 241:1077–1080 (1998).

Lee AS, Gutiérrez-Arcelus M, Perry GH, Vallender EJ, Johnson WE, et al: Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. Hum Mol Genet 17: 1127–1136 (2008).

Lee HH, Chang SF, Tseng YT, Lee YJ: Identification of the size and antigenic determinants of the human *C4* gene by a polymerase chain-reaction-based amplification method. Anal Biochem 357:122–127 (2006).

Linzmeier RM, Ganz T: Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22–p23. Genomics 86:423–430 (2005).

Livak KJ, Flood SJ, Marmaro J, Giusti W, Deetz K: Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. PCR Methods Appl 4:357–362 (1995).

Marklund S, Kijas J, Rodriguez-Martinez H, Rönnstrand L, Funa K, et al: Molecular basis for the dominant white phenotype in the domestic pig. Genome Res 8:826–833 (1998).

Morrison TB, Weis JJ, Wittwer CT: Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques 24:954–962 (1998).

Müller B, Zöpf K, Bachofer J, Steimer W: Optimized strategy for rapid cytochrome P450 2D6 genotyping by real-time long PCR. Clin Chem 49: 1624–1631 (2003).

Nakajima T, Ohtani H, Naruse T, Shibata H, Mimaya JI, et al: Copy number variations of *CCL3L1* and long-term prognosis of HIV-1 infection in asymptomatic HIV-infected Japanese with hemophilia. Immunogenetics 59: 793–798 (2007).

Nevilie M, Selzer R, Aizenstein B, Maguire M, Hogan K, et al: Characterization of cytochrome P450 2D6 alleles using the Invader system. Biotechniques Suppl:34–43 (2002).

Park CH, Jeong HJ, Choi YH, Kim SC, Jeong HC, et al: Systematic analysis of cDNA microarray-based CGH. Int J Mol Med 17:261–267 (2006).

Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, et al: Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci USA 103:8006–8011 (2006).

Pielberg G, Olsson C, Syvänen AC, Andersson L: Unexpectedly high allelic diversity at the *KIT* locus causing dominant white color in the domestic pig. Genetics 160:305–311 (2002).

Pielberg G, Day AE, Plastow GS, Andersson L: A sensitive method for detecting variation in copy numbers of duplicated genes. Genome Res 13:2171–2177 (2003).

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207–211 (1998).

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41–46 (1999).

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al: Global variation in copy number in the human genome. Nature 444:444–454 (2006).

Ronaghi M, Pettersson B, Uhlén M, Nyrén P: PCR-introduced loop structure as primer in DNA sequencing. Biotechniques 25: 876–884 (1998).

Schaeffeler E, Schwab M, Eichelbaum M, Zanger UM: CYP2D6 genotyping strategy based on gene copy number determination by TaqMan real-time PCR. Hum Mutat 22:476–485 (2003).

Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470 (1995).

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

Seo BY, Park EW, Ahn SJ, Lee SH, Kim JH, et al: An accurate method for quantifying and analyzing copy number variation in porcine *KIT* by an oligonucleotide ligation assay. BMC Genet 8:81 (2007).

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, et al: Segmental duplications and copy number variation in the human genome. Am J Hum Genet 77:78–88 (2005).

She X, Liu G, Ventura M, Zhao S, Misceo D, et al: A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res 16: 576–583 (2006)

Shen F, Huang J, Fitch KR, Troung VB, Kirby A, et al: Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. BMC Genet 9:27 (2008).

Trask BJ, Friedman C, Martin-Gallardo A, Rowen L, Akinbami C, et al: Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Hum Mol Genet 7:13–26 (1998).

Wong KK, Deleeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al: A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104 (2007).

Wu DY, Wallace RB: Specificity of the nick-closing activity of bacteriophage T4 DNA ligase. Gene 76:245–254 (1989).

Wu YL, Savelli SL, Yang Y, Zhou B, Rovin BH, et al: Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs in 50 consanguineous subjects with defined HLA genotypes. J Immunol 179:3012–3025 (2007).

Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, et al: 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 (2007).

Zackrisson AL, Lindblom B: Identification of CYP2D6 alleles by single nucleotide polymorphism analysis using pyrosequencing. Eur J Clin Pharmacol 59:521–526 (2003).