Communications of the Korean Statistical Society
2009, Vol. 16, No. 6, 1037–1046

Comparison of Methods for Detecting and Quantifying Variation in Copy Numbers of Duplicated Genes

Jin-Tae Jeon[a], Sung Jin Ahn[1,b]

[a]Division of Animal Life Science, Gyeongsang National University
[b]Department of Information Statistics, RICIC and RINS, Gyeongsang National University

Abstract

Copy number variations(CNVs) are known as one of the most important factors in susceptibility to genetic disorders because they affect expression levels of genes. In previous studies, pyrosequencing, mini-sequencing, real-time polymerase chain reaction(PCR), invader assays and other techniques have been used to detect CNVs. However, the higher the copy number in a genome, the more difficult it is to resolve the copies, so a more accurate method for measuring CNVs and assigning genotype is needed. PCR followed by a quantitative oligonucleotide ligation assay(qOLA) was developed for quantifying CNVs. The aim of this study was to compare the two methods for detecting and quantifying the CNVs of duplicated gene: the published pyrosequencing assay(pyro_CNV) and the newly developed qOLA_CNV. The accuracy and precision of the assay were evaluated for porcine KIT, which was selected as a model locus. Overall, the root mean squares(RMSs) of bias and standard deviation of qOLA_CNV were 2.09 and 0.45, respectively. These values are less than half of those of pyro_CNV.

Keywords: Copy number variation, quantitative oligonucleotide ligation assay, pyrosequencing assay, root mean square.

1. Background

Genetic disorder susceptibility is known to be associated with genetic variation such as single nucleotide polymorphisms(SNPs) and structural variation including copy number variations(CNVs) (Redon *et al.*, 2006; Stranger *et al.*, 2007; Kehrer-Sawatzki, 2007). Therefore, once identified, a CNV needs to be analyzed at the locus level, and ultimately, the genotype and haplotype must be determined to elucidate its relationship with a particular genetic alteration. Pyrosequencing, mini-sequencing, real-time PCR and invader assays are among the techniques that have been used to detect CNVs (Pielberg *et al.*, 2003; Nevilie *et al.*, 2002; Aldred *et al.*, 2005).

The porcine KIT was selected for this study because it is a well characterized and functionally important CNV. The Dominant White/KIT locus that determines white coat color is located in Sus scrofa chromosome 8(SSC8) (Johansson, 1996; Hirooka *et al.*, 2002). Two KIT mutations cause the Dominant White phenotype in pigs: a gene duplication associated with a partially dominant phenotype, which is depicted as normal and duplicated in Figures 1(a) and (b), and a splice mutation leading to the fully dominant allele (Johansson, 1996; Marklund, 1998), which is marked in Figure 1(a) as an SNP(G/A) at the first nucleotide of intron 17.

As shown in Figure 1(c), there are four known major alleles at the KIT locus: 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 (Johansson *et al.*, 1992; Giuffra *et al.*, 1999). I allele

[1] Corresponding author: Professor, Department of Information Statistics, RICIC and RINS, Gyeongsang National University, Gyeongnam 660-701, Korea. E-mail: ahnsj@gnu.ac.kr

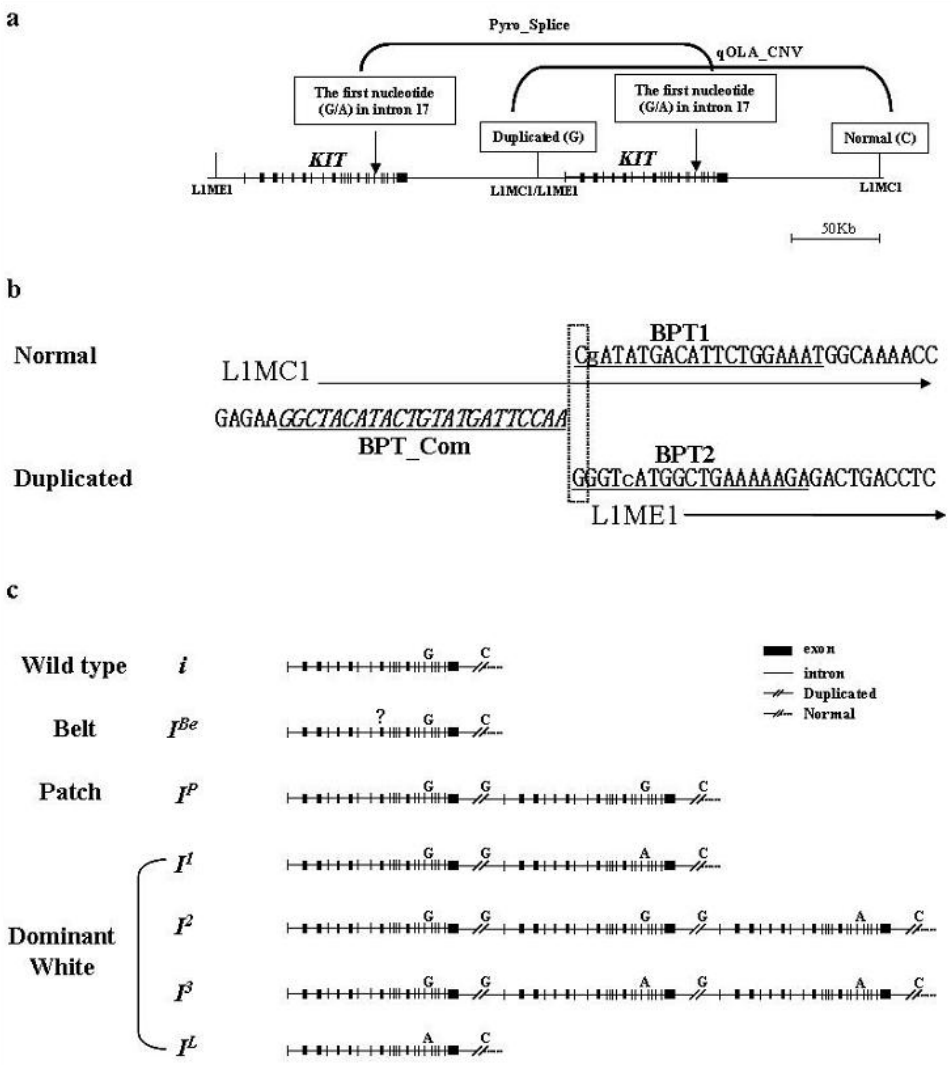


Figure 1: *(a) A schematic description of tandem duplication at the porcine KIT locus; (b) Nucleotide sequence around the breakpoint; (c) Schematic descriptions of KIT alleles*

diversity has been reported and classified in detail as I^1, I^2, I^3 and I^L (Pielberg *et al.*, 2003). All possible genotypes, which are derived from the alleles shown in Figure 1(c), and theoretical ratios of spliced and duplicated copies corresponding to each genotype are presented in Table 1. The two ratios of each polymorphism were used as reference values when the genotypes of experimental pig samples were assigned in this study.

To analyze the KIT locus, RFLP (Marklund *et al.*, 1998), minisequencing, real-time PCR (Pielberg *et al.*, 2002), invader and pyrosequencing assays (Pielberg *et al.*, 2003) have been used. Pyrosequencing has provided the best resolution for quantifying KIT CNV giving more accurate results than real-time PCR amplification and invader technologies. However, as the copy number increases, it

Table 1: Theoretical genotype description of the KIT locus by the splice mutation and copy number variation

Genotype[a]	Spliced copy to Total copy	Ratio of [b] spliced(%)	Ratio of [c] duplicated(%)	Seed Number [d]
$i/i(I^{Be})$	0 : 2	0	0	1
$I^P/i(I^{Be})$	0 : 3	0	33.3	2
I^P/I^P	0 : 4	0	50	3
I^2/I^P	1 : 5	20	60	4
I^1/I^P	1 : 4	25	50	5
$I^2/i(I^{Be})$	1 : 4	25	50	5
$I^2/i(I^{Be})$	1 : 3	33.3	33.3	6
I^2/I^2	2 : 6	33.3	66.7	7
I^1/I^2	2 : 5	40	60	8
I^3/I^P	2 : 5	40	60	8
I^1/I^1	2 : 4	50	50	9
$I^3/i(I^{Be})$	2 : 4	50	50	9
I^2/I^3	3 : 6	50	66.7	10
I^1/I^3	3 : 5	60	60	11
I^3/I^3	4 : 6	66.7	66.7	12

a: I^L allele is not included because it is a very rare allele that has been reported once in a synthetic line by crossing Large White and Meishan breeds (Pielberg *et al.*, 2003).
b: This is the reference ratio for pyro_Splice.
c: This is the reference ratio for qOLA_CNV and pyro_CNV. Duplicated copy number = Total copy number − 2.
d: These are numbers of class centroids used for nearest centroid sorting.

gradually becomes more difficult to use the pyrosequencing method to accurately distinguish among genotype classes that differ by only one copy. This is because the relative increase in the signal from the duplicate breakpoint becomes smaller (Pielberg *et al.*, 2003). An underestimated CNV ratio may result in an ambiguous genotype assignment in samples for which family information, including parental genotypes, is not available.

We have therefore developed PCR followed by a quantitative oligonucleotide ligation assay(qOLA) which gives high resolution data for determining KIT CNV, especially if the copy number is high (> 4). The development of qOLA is based on the strategy previously described in Pielberg *et al.* (2003), but it improves on the pyrosequencing method (Pielberg *et al.*, 2003) for analyzing CNV of the locus. We have also established a nearest centroid sorting procedure to verify the reliability of the genotype assignment for random animal samples. The qOLA used on a platform with an ABI sequencer is sensitive enough to analyze DNA from a few hair follicles, so DNA from various sources could be used for qOLA.

2. Results

2.1. Verifying the specificity of the PCR primers used for analyzing KIT CNV

The PCR primers designed for the published pyrosequencing method (Pielberg *et al.*, 2003) were used in this study. The primer sequences selected from the KIT duplication breakpoint are located on repetitive elements, L1MC1 and L1ME1 (Figure 1(a) and (b)). The forward primer (KITBPF) shows 80% sequence identity with the L1MC1 consensus sequence and the two reverse primers, KIT1BPR for the normal copy and KIT2BPR for the duplicated copy, show 63.2% and 94.7% sequence identity with L1MC1 and L1ME1, respectively. This finding raised the question of whether the PCR products may contain nonspecific amplification products from other genomic regions. To evaluate the specificity of the PCR primers, somatic cell hybrid panel mapping was performed prior to the quantification assay. The two amplicons were located in SSC8p11, where the KIT locus exists (assignment prob-

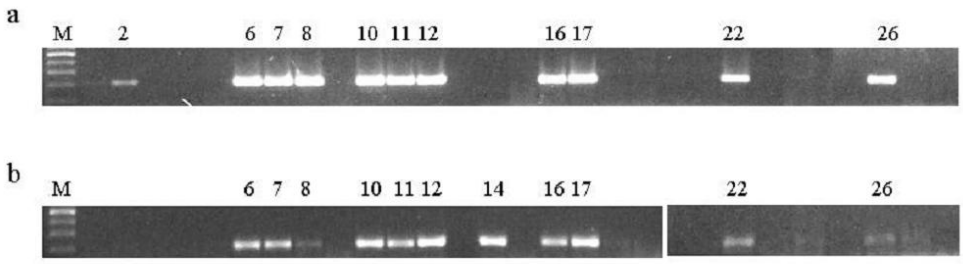
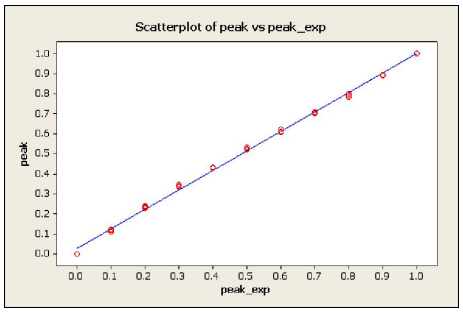
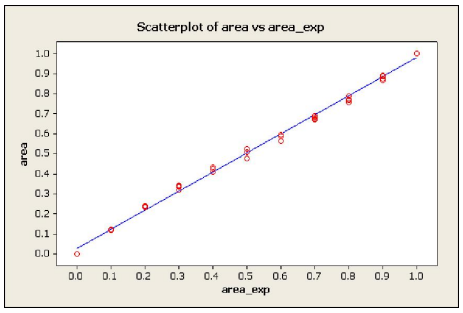


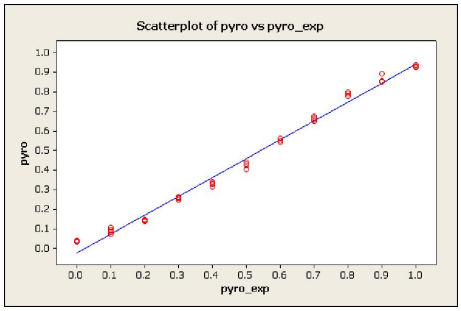
Figure 2: *(a) A schematic description of tandem duplication at the porcine KIT locus; (b) Nucleotide sequence around the breakpoint*



(a) reference *vs.* peak



(b) reference *vs.* area



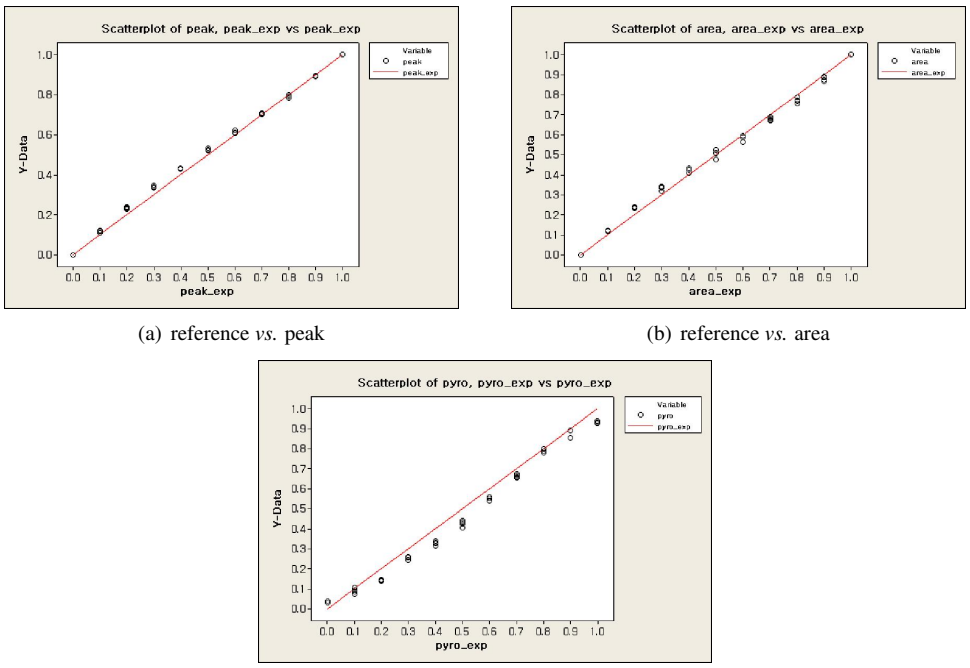
(c) reference *vs.* pyro

Figure 3: *Standard curves (or Regression plots): Expected ratio (reference) of duplicated copies (horizontal axis) vs. (a) Ratio of peak height in qOLA_CNV; (b) Ratio of peak area in qOLA_CNV; (c) Ratio in pyro_CNV*

ability/correlation: 0.8789/0.9250 for normal and 0.8791/0.9250 for duplicated), indicating that the amplifications of the primer sets were specific. As shown in Figure 2, the primer sets were clearly amplified.

2.2. Evaluation of the established qOLA to measure the CNV of KIT (qOLA_CNV)

The amplicons of the duplicated and normal copies were cloned into the pCR®2.1-TOPO vector (Invitrogen, USA). The cloned amplicons were re-amplified using the M13 forward and reverse primers, and were then purified and serially diluted from 0% to 100% duplicated copy *vs.* normal copy. PCR followed by qOLA_CNV was performed on four replicates and two standard curves were obtained for



(a) reference *vs.* peak

(b) reference *vs.* area

(c) reference *vs.* pyro

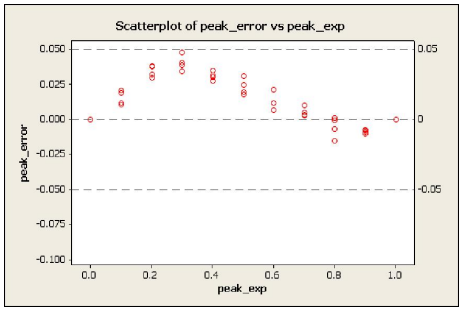
Figure 4: *Calibration plots: Expected ratio (reference) of duplicated copies (horizontal axis) vs. (a) Ratio of peak height in qOLA_CNV; (b) Ratio of peak area in qOLA_CNV; (c) Ratio in pyro_CNV*

peak height (abbreviated as "peak" in the plots) and peak area (abbr. "area"), as depicted in Figure 3(a) and (b). qOLA_CNV was compared with the published pyrosequencing assay (Pielberg *et al.*, 2003) for KIT CNV (pyro_CNV) (abbr. "pyro"). The same serial dilutions used for qOLA_CNV were used to obtain the standard curve for pyro_CNV, as depicted in Figure 3(c).

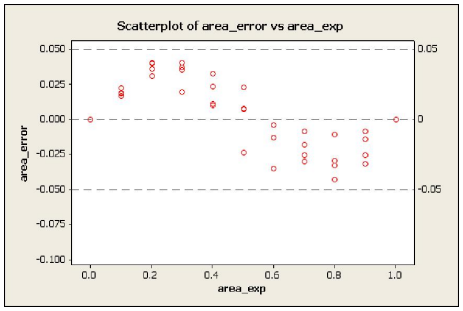
All the three standard curves or regression lines in Figure 3 seemed to show almost the same linearities, indicating almost the same performance concerning the type and magnitude of errors. Actually correlation coefficients were 0.999 for both standard curves in qOLA_CNV and 0.995 (0.997 in Pielberg *et al.*, 2003) in the standard curve of pyro_CNV. The correlation coefficient is simply an index of the linearity of the standard curve or the regression line. To examine the hidden patterns of errors, however, we ought to compare data values to the calibration lines of the reference values, rather than to the regression lines of the fitted values. Calibration plots in Figure 4 display the data values with the calibration lines and reveal part of the real performances of the three methods, as was expected. Error plots in Figure 5 highlight the magnitudes and the nonlinear patterns of errors.

Now we need to evaluate the three methods in terms of precision and accuracy. Precision refers to random errors, whereas accuracy refers to systematic errors (Ahn, 2007; Westgard and Hunt, 1973). We will measure accuracy using the bias, which is the difference between the mean of the replicates and the reference point, and measure precision using the standard deviation(SD) of replicates for a reference point,

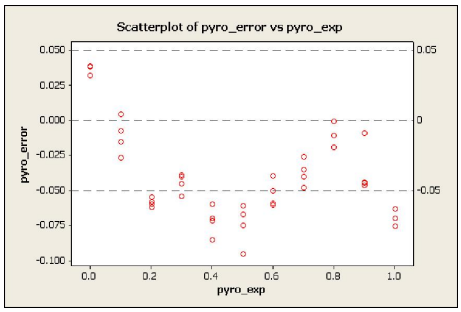
As shown in Table 2 and Bias plots in Figure 6 and SD plots id Figure 7, peak height values in qOLA_CNV fit the reference values better and show the least variation. In particular, for accurate genotyping of individuals with a total of more than 4 KIT copies, the assay needs better resolution in



(a) reference *vs.* peak_error



(b) reference *vs.* area_error



(c) reference *vs.* pyro_error

Figure 5: *Error plots for (a) ratio of peak height in qOLA_CNV; (b) ratio of peak area in qOLA_CNV; (c) ratio in pyro_CNV*

Table 2: Bias and SD of each method for each duplicated ratio

Duplicated copy ratio (%)	Bias of qOLA_CNV (Height, %)	SD of qOLA_CNV (Height, %)	Bias of qOLA_CNV (Area, %)	SD of qOLA_CNV (Area, %)	Bias of pyro_CNV (%)	SD of pyro_CNV (%)
0	0.00	0.00	0.00	0.00	3.64	0.38
10	1.55	0.51	1.91	0.24	−1.13	1.30
20	3.43	0.42	3.67	0.46	−5.87	0.31
30	4.01	0.56	3.31	0.93	−4.48	0.69
40	3.09	0.31	1.92	1.08	−7.17	1.04
50	2.32	0.58	3.40	1.95	−7.48	1.49
60	1.27	0.61	−1.65	1.32	−5.23	0.97
70	0.51	0.33	−2.07	0.93	−3.75	0.93
80	−0.56	0.72	−2.91	1.34	−1.04	0.92
90	−0.87	0.13	−2.01	1.03	−3.62	1.81
100	0.00	0.00	0.00	0.00	6.98	0.52
RMS1[a]	2.09	0.45	2.16	1.03	5.05	1.04
RMS2[b]	0.85	0.51	2.21	1.17	3.73	1.21

a: Overall RMS
b: RMS for the zone between 60–90%

the zone between 60% and 90% in the standard curve. We can summarizes the sizes of biases and SDs at 10 reference points by RMS(root mean squares).

In this zone, the peak area values in qOLA_CNV showed RMSs of the bias and SD as 2.21 and 1.17, respectively. In contrast, the RMSs of the bias and SD of the peak height values in qOLA_CNV

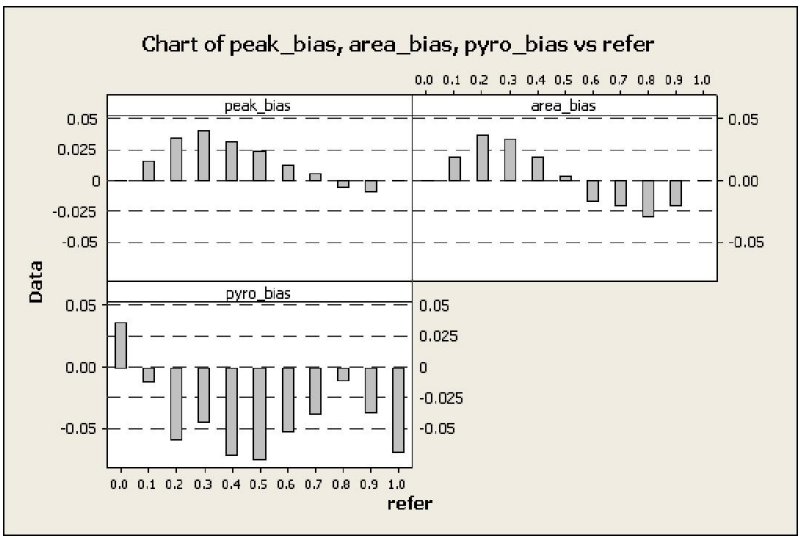


Figure 6: *Bias plots for (a) ratio of peak height in qOLA_CNV; (b) ratio of peak area in qOLA_CNV; (c) ratio in pyro_CNV*

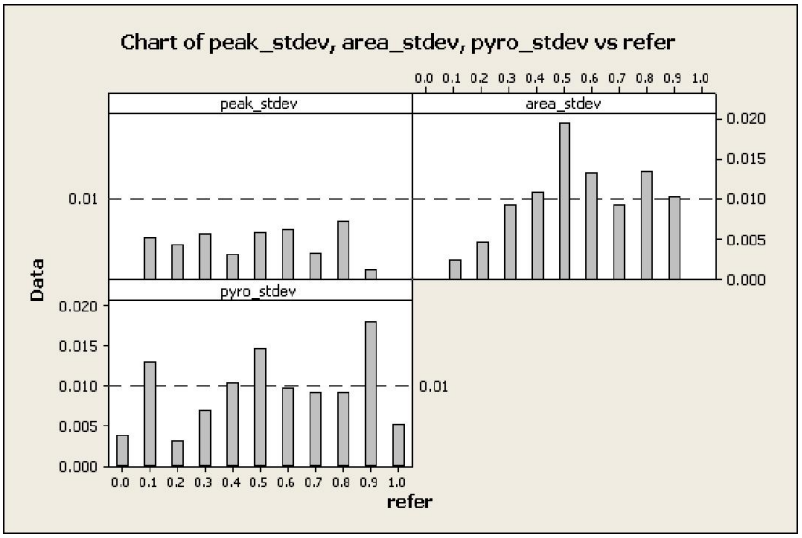


Figure 7: *SD plots for (a) ratio of peak height in qOLA_CNV; (b) ratio of peak area in qOLA_CNV; (c) ratio in pyro_CNV*

were 0.86 and 0.51, respectively, in the same zone. The overall RMSs of the bias (5.05) and SD (1.04) in pyro_CNV were more than twice those for the peak height in qOLA_CNV (2.09 and 0.45).

The RMS plot displays the RMSs of biases and SDs for the three methods in Figure 8. In conclusion, CNV estimation for porcine KIT using the peak height values in qOLA_CNV showed the lowest systematic errors and variations of the studied methods, and therefore was used in further experiments to analyze KIT CNV and assign genotypes.

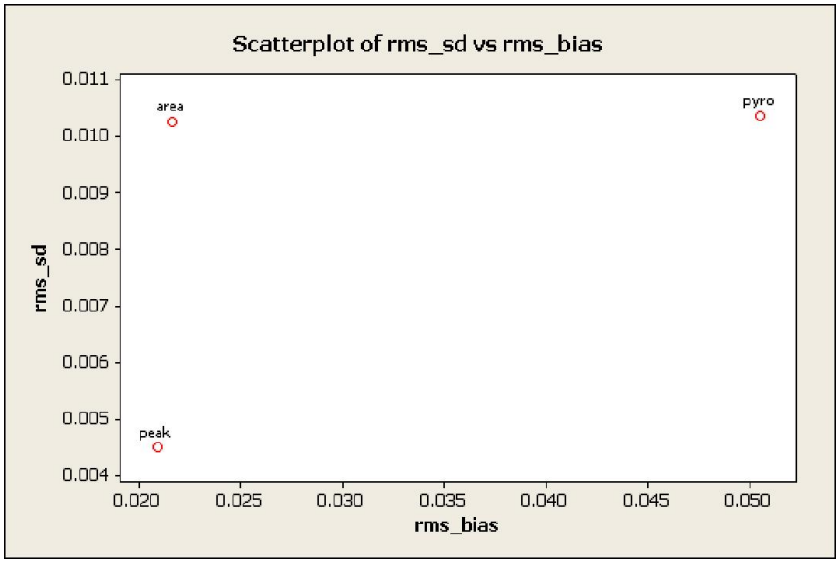


Figure 8: *RMS plot of biases vs. SDs for (a) ratio of peak height in qOLA_CNV; (b) ratio of peak area in qOLA_CNV; (c) ratio in pyro_CNV*

2.3. Application of the proposed method to CNV and genotyping tests

Another pyrosequencing assay pyro_Splice (Pielberg *et al.*, 2002) for quantifying KIT copies with a splicedonor mutation in intron 17 was combined with qOLA_CNV. qOLA_CNV gives information about the total KIT copy numbers in a sample, and pyro_Splice gives additional information about the ratio of spliced KIT copies to the total copies estimated by qOLA_CNV. As shown in Table 1, several different genotypes have identical ratios in each polymorphism. Therefore, combined information from the two polymorphisms should yield better discriminating power in assigning genotype.

CNV and genotyping tests using a combination of qOLA_CNV and pyro_Splice were performed to verify KIT allele segregation. For genotype assignment we used two classification methods based on the clusters of measurements on a scatter plot and the clusters of observations at 12 seed points using nearest centroid sorting (Everitt, 1974) implemented in PROC FASTCLUS of the SAS 9.1 package (SAS, 2004). The genotyping results showed 100% agreement between the two methods and were in good agreement with both the theoretical genotype ratios and phenotypes. The detailed procedures and results were presented in Seo *et al.* (2007).

3. Conclusion

We have compared the two methods for detecting and quantifying the CNVs of duplicated gene: the published pyrosequencing assay(pyro_CNV) and the newly developed qOLA_CNV. The accuracy and precision of the assay were evaluated for porcine KIT, which was selected as a model locus. CNV estimation for porcine KIT using the peak height values in qOLA_CNV showed the lowest systematic errors and variations among the studied three procedures: the peak-height qOLA_CNV, the peak-area qOLA_CNV, and the pyro_CNV. And therefore the peak-height qOLA_CNV was used in further experiments to analyze KIT CNV and assign genotypes. Further, we have established a reliable assay for measuring tandem CNV that could be applied for a variety of samples, such as those in a known

pedigree, those with predictable segregation, those without pedigree information, and genomic DNA of poor quality. Combining this method with a verification procedure using statistical clustering, genotypes can be successfully assigned with high confidence. This development could be widely applicable to studies of the function and mechanism of CNV in other species, and may be particularly useful for tandemly repeated CNV.

References

Ahn, S. J. (2007). *Statistical Quality Control using Minitab 14*, Free Academy, Seoul, 260–262.

Aldred, P. M., Hollox, E. J. and Armour, J. A. (2005). Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3, *Human Molecular Genetics*, **14**, 2045–2052.

Everitt, B. (1974). *Cluster Analysis*, Heinemann Educational Books, London.

Giuffra, E., Evans, G., Törnsten, A., Wales, R., Day, A., Looft, H., Plastow, G. and Andersson, L. (1999). The Belt mutation in pigs is an allele at the Dominant White (I/KIT) locus, *Mammalian Genome*, **10**, 1132–1136.

Hirooka, H., de Koning, D. J., van Arendonk, J. A. M., Harlizius, B., de Groot, P. N. and Bovenhuis, H. (2002). Genome scan reveals new coat color loci in exotic pig cross, *Journal of Heredity*, **93**, 1–8.

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

Johansson, M., Moller, M., Ellegren, H., Marklund, L., Gustavsson, U., Ringmar-Cederberg, E., Andersson, K., Edfors-Lilja, I. and Andersson, L. (1992). The gene for dominant white color in the pig is closely linked to ALB and PDGRFRA on chromosome 8, *Genomics*, **14**, 965–969.

Kehrer-Sawatzki, H. (2007). What a difference copy number variation makes, *BioEssays*, **29**, 311–313.

Marklund, S., Kijas, J., Rodriguez-Martinez, H., Rönnstrand, L., Funa, K., Moller, M., Lange, D., Edfors-Lilja, I. and, Andersson, L. (1998). Molecular basis for the dominant white phenotype in the domestic pig, *Genome Research*, **8**, 826–833.

Nevilie, M., Selzer, R., Aizenstein, B., Maguire, M., Hogan, K., Walton, R., Welsh, K., Neri, B. and de Arruda, M. (2002). Characterization of cytochrome P450 2D6 alleles using the Invader system, *Biotechniques*, Suppl 34–38, 40–43.

Pielberg, G., Day, A. E., Plastow, G. S. and Andersson, L. (2003). A sensitive method for detecting variation in copy numbers of duplicated genes, *Genome Research*, **13**, 2171–2177.

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

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. and Hurles, M. E. (2006). Global variation in copy number in the human genome, *Nature*, **444**, 444–454.

SAS Institute Inc. (2004). SAS/STAT 9.1 user's guide Volume 6. SAS Institute, Cary, N.C., 1377–1427.

Seo, B. Y., Park, E. W., Ahn, S. J., Lee, S. H., Kim, J. H., Im, H. T., Lee, J. H., Cho, I. C., Kong, I. K. and Jeon, J. T. (2007). An accurate method for quantifying and analyzing copy number variation in porcine KIT by an oligonucleotide ligation assay, *BMC Genetics*, **8**, 81.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi A, Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E. and Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes, *Science*, **315**, 848–853.

Westgard, J. O. and Hunt, M. R. (1973). Use and interpretation of common statistical tests in method comparison studies, *Clinical Chemistry*, **19**, 49–57.

Received August 2009; Accepted October 2009