Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Positional effects of polymorphisms in probe-target sequences on genoplot images of oligonucleotide microarrays

T.L. Cui[1], H. Nakaoka[1], K. Akiyama[1], H. Kamura[1], K. Hosomichi[1], J. Bae[2], H. Cheong[2], H. Shin[2,3], T. Yada[4] and I. Inoue[1]

[1]Division of Molecular Life Science, School of Medicine,
Tokai University, Kanagawa, Japan
[2]Department of Life Science, Sogang University, Seoul, Korea
[3]Department of Genetic Epidemiology, SNP Genetics, Inc., Complex B, Seoul, Korea
[4]Graduate School of informatics, Kyoto University, Kyoto, Japan

Corresponding author: T.L. Cui
E-mail: tailinc@is.icc.u-tokai.ac.jp

Genet. Mol. Res. 9 (1): 524-531 (2010)
Received November 20, 2009
Accepted January 4, 2010
Published March 23, 2010

ABSTRACT. Single nucleotide polymorphisms (SNPs) present in probe-target sequences (SPTS) have been shown to be associated with abnormal genoplot images. We explored the effects of SPTS positions on genoplot images using a data set from a genome-wide association study typed on an Illumina Human Hap300 platform. We screened the physical genomic positions of 308,330 autosomal probes to identify SPTS candidates deposited in dbSNP. The genoplot images across 293 individuals were inspected further in SNPs bearing an SPTS candidate. We identified 35,185 SNPs bearing a single SPTS candidate, including 264 SNPs showing abnormal genoplot images. The frequencies of SPTS at distances within 10 bases from the target SNP were significantly higher in the 264 SNPs showing abnormal genoplot images, than in the remaining 34,921 SNPs (49.62 *vs* 12.87%; Fisher exact test; $P = 2.2 \times 10^{-16}$). Of these 264 SNPs, we randomly selected 20 SNPs and resequenced them in 97 individuals. An SPTS within 10 bases of the target SNP was confirmed in all 20 SNPs, except for one SNP with a small deletion (7 bases) in the probe-target sequence. Taken together, these results suggest an association of a proximal SPTS with an abnormal genoplot image, which could result in spurious genotype detections, highlighting the importance of minimizing systematic errors in microarray experiments.

Key words: Probe-target sequence; Genoplot image; Positional effects; Oligonucleotide microarray

INTRODUCTION

A typical genoplot image of a genotyping microarray shows three discrete groups corresponding to homozygous and heterozygous genotypes (Gunderson et al., 2005). However, in some situations, the genoplot represents an abnormal cluster pattern (abnormal genoplot image), which interferes with genotype detections (Franke et al., 2008).

An abnormal genoplot image may represent a biologically meaningful failure, for instance, the presence of untyped third alleles, such as a deletion copy number variation (CNV) mapped within a single nucleotide polymorphism (SNP) site, or a polymorphism present in probe-target sequences (SPTS) (McCarroll et al., 2006; Franke et al., 2008). In case of a deletion CNV, the abnormal genoplot image is shown to correspond to individuals with hemizygous and homozygous deletions. Consequently, individual genotypes with a hemizygous deletion are miscalled as homozygotes, while individual genotypes with homozygous deletions are called as missing (McCarroll et al., 2006).

Hybridization-based technologies, such as microarray, rely on the precise probe interaction between a microarray probe and its target sequence to ensure specific and accurate signal intensity measurements (Fan et al., 2007; Sliwerska et al., 2007; Zhang et al., 2007; Bemmo et al., 2008; Wang et al., 2008). The presence of SNP present in an SPTS may affect hybridization affinities due to single nucleotide mismatch of SPTS with a microarray probe, consequently leading to an abnormal genoplot image and false-positive results (Franke et al., 2008; Benovoy et al., 2008).

With the increasing volume of SNPs deposited in public databases, such as NCBI dbSNP, it is possible that there is an SNP within the immediately adjacent locus that is complementary to the 50-base probe (used in the Illumina Infinium chemistry). This raises questions of which positions of SPTS are associated with the abnormal genoplot image.

In the present study, we explored the association of SPTS positions with the genoplot image, using a data set from a genome-wide association study typed on Illumina Human Hap300 platform.

MATERIAL AND METHODS

Data sets for SPTS discovery

A data set from our unpublished genome-wide association study (GWAS) of 293 unrelated individuals that passed quality control and that had been typed on the Illumina Infinium II Human Hap300 BeadChip platform for 317,503 SNPs was used in this analysis. These included 97 controls and 196 patients with intracranial aneurysm in a Japanese population. The

Ethics Committees of Tokyo Women's Medical University approved the study protocols, and all participants gave written informed consent.

Identification of SNPs bearing SPTS candidate deposited in dbSNP

To identify these SPTS candidates, we screened the physical genomic positions in which the 50-base probes annealed and determined whether more SNPs were known in these loci in dbSNP (Build 127) using an automated in-house developed algorithm. According to this algorithm, a BLAST hit is considered to be an SPTS candidate if an SNP of dbSNP is located within a probe-targeted genomic region. All analyses were performed on the NCBI Build 36 Genomic Assembly. We identified 35,185 SNPs bearing a single SPTS candidate (fraction D in Figure 1) and 5317 SNPs bearing multi-SPTS (at least 2 SPTS candidates) (fraction C in Figure 1). Sets of 5317 SNPs and 15,139 SNPs that mapped within a common CNV (individual frequency above 5%) deposited in Database of Genomic Variants (http://projects.tcag.ca/variation) were removed from our assay (fraction B in Figure 1).

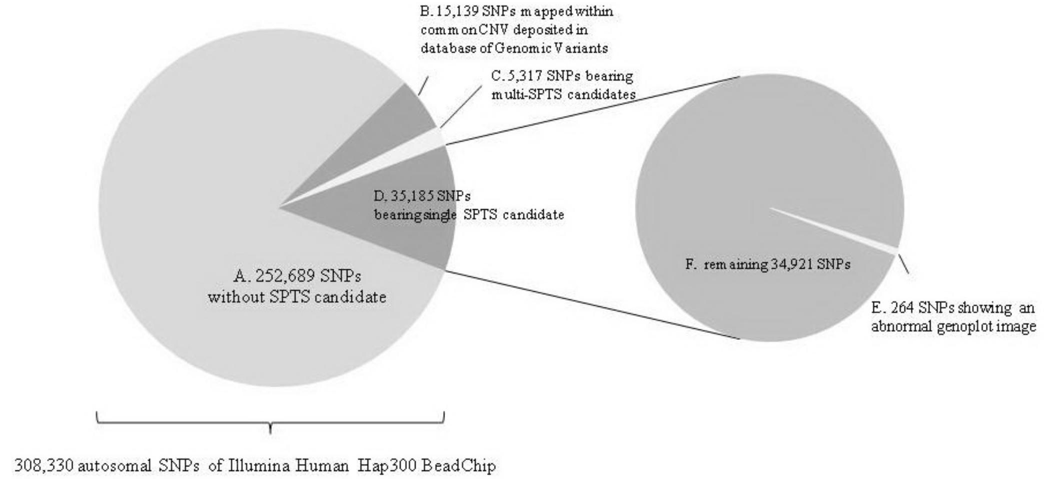
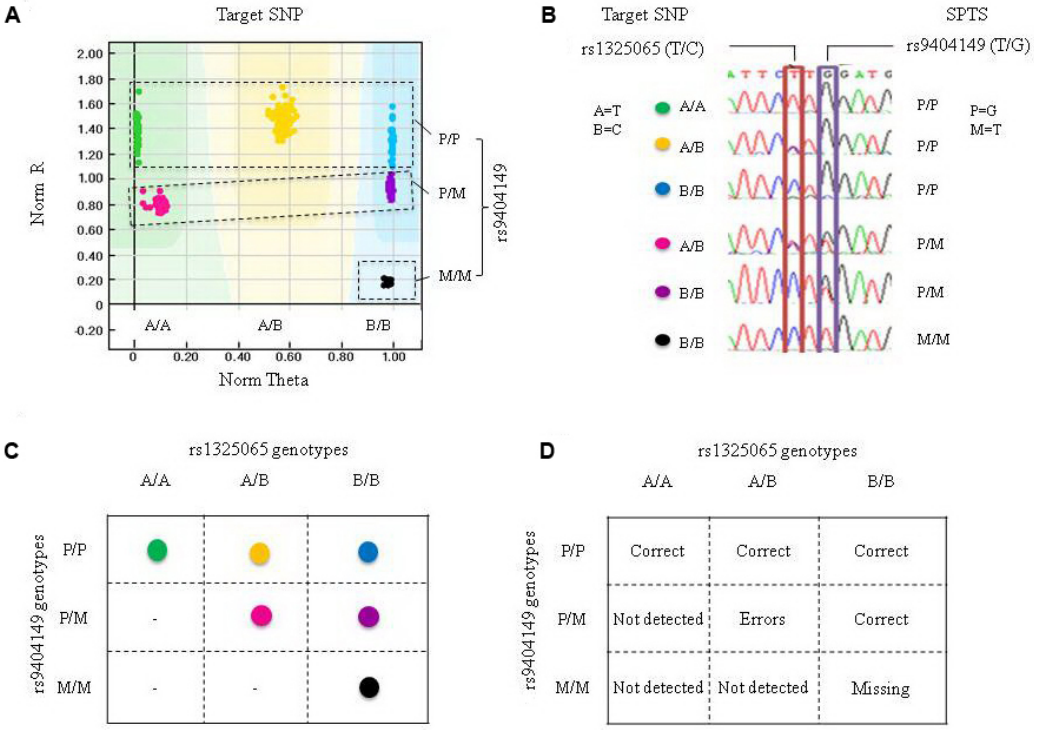


Figure 1. Identification of single nucleotide polymorphism (SNP) candidates with the presense of probe-target sequence (SPTS) of Illumina Human Hap300 BeadChip. **A.** 252,689 SNPs without SPTS candidate. **B.** 15,139 SNPs mapped within common copy number variation (CNV) deposited in Database of Genomic Variants. **C.** 5,317 SNPs bearing multiple SPTS candidates. **D.** 35,185 SNPs bearing single SPTS candidate. **E.** 264 SNPs showing an abnormal genoplot image. **F.** Remaining 34,921 SNPs bearing single SPTS candidate.

Visual inspection of genoplot

Typically, the genoplot image shows three typical clusters (A/A, A/B and B/B) corresponding to homozygous and heterozygous genotypes, when the minor allele frequency is sufficiently high (Sapolsky et al., 1999; Chen and Kwok, 1999; van Heel et al., 2007). However, the individual genotypes were grouped into additional two or three extra-clusters that fall below expectations in the presence of an SPTS (Franke et al., 2008). We refer to this as an abnormal genoplot image. For instance, individual genotypes of rs1325065 (T/C) with an SPTS and rs9404149 (T/G) positioned at 2 bases from rs1325065 showed an abnormal genoplot image consisting of three typical clusters (green, yellow and blue circles) and three extra-clusters (pink, violet and black circles) (Figure 2A).



	A/A	A/B	B/B
P/P	●	●	●
P/M	-	●	●
M/M	-	-	●

	A/A	A/B	B/B
P/P	Correct	Correct	Correct
P/M	Not detected	Errors	Correct
M/M	Not detected	Not detected	Missing

Figure 2. Abnormal genoplot image corresponding to the presence of single nucleotide polymorphism (SNP) present in a probe-target sequence (SPTS). **A.** Genoplot of rs1325065 (T/C) bearing an SPTS, rs9404149 (T/G) positioned at distances 2 bases from rs1325065. Individual genotypes grouped into six clusters comprising three typical clusters (A/A: green, A/B: yellow, and B/B: blue circles) and three extra-clusters (A/B: pink, B/B: violet, and B/B: black circles). P/P: perfect match allelic homozygote of rs9404149, P/M: heterozygote of rs9404149, and M/M: mismatch allelic homozygote of rs9404149. The clusters included in the dashed frames are shown to correspond to the genotypes P/P, P/M and M/M of rs9404149, respectively. **B.** Resequencing for region flanking rs1325065 and rs9404149. The color-coded circles represent the genotypes of both rs1325065 and rs9404149, corresponding to the clusters in *Panel A*. **C.** Genotype combination between rs1325065 (target SNP) and rs9404149 (SPTS) for each cluster. **D.** Genotype combinations between rs1325065 (target SNP) and rs9404149 (SPTS) for missing and erroneous genotypes.

To identify SNPs showing an abnormal genoplot image, we carried out visual inspection of the genoplot image in these 35,185 SNPs bearing a single SPTS candidate using the Beadstudio software (Illumina® Beadstudio 3.0), and observed 264 SNPs showing an abnormal genoplot image (group E in Figure 1).

Resequencing

Of these 264 SNPs showing an abnormal genoplot image, we randomly selected 20 SNP candidates with an SPTS, and carried out direct sequencing in 97 control individuals using the ABI 3130XL Genetic Analyzer sequencer (Applied Biosystems).

RESULTS AND DISCUSSION

In this study, we identified 35,185 SNPs bearing a single SPTS candidate (fraction D in Figure 1), of which 264 SNPs showed an abnormal genoplot image (fraction E in Figure 1). We found that the frequencies of SPTS positioned at distances ranging from 1 to 50 bases (probe size) gradually decrease with proximity to the target SNP in 35,185 SNPs (Figure 3A), which is apparently different from expected uniform distributions (Zhao and Boerwinkle, 2002; Madsen et al., 2007). However, their frequency at distance 1 base is over-represented, because of the high CpG mutation rate (Hwang and Green, 2004).

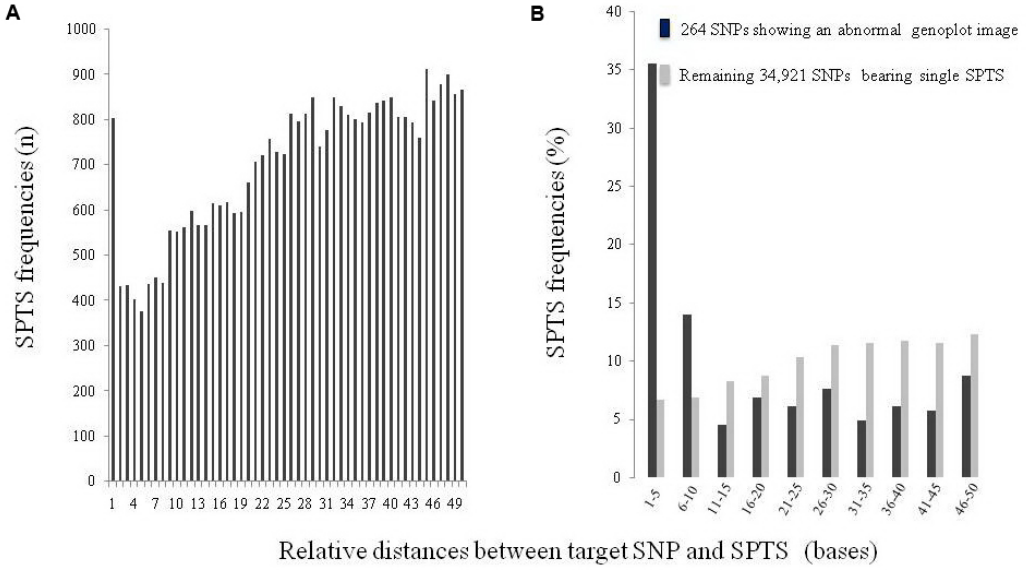


Figure 3. Distributions of probe-target sequence (SPTS) frequencies at different distances ranging from 1 to 50 bases relative to target single nucleotide polymorphism (SNP). **A.** SPTS frequencies in 35,185 SNPs bearing single SPTS candidate. **B.** SPTS frequencies in 264 SNPs showing an abnormal genoplot image and remaining 34,921 SNPs.

The frequencies of SPTS positioned at distances within 10 bases (defined as proximal SPTS) from the target SNP are significantly higher in these 264 SNPs showing an abnormal genoplot image, than in the remaining 34,921 SNPs without an abnormal genoplot image (49.62 *vs* 12.87%; Fisher exact test; $P = 2.2 \times 10^{-16}$) (Figure 3B). In contrast, the frequencies of SPTS positioned at distances 11 to 50 bases (defined as distal SPTS) are significantly higher in 34,921 SNPs than the 264 SNPs showing an abnormal cluster pattern. This different distribution of SPTS frequencies between SNPs with and without an abnormal genoplot image suggests the association of proximal SPTS with the abnormal genoplot image.

Of these 264 SNPs, we randomly selected 20 SNPs and resequenced them in 97 individuals to confirm the presence of a proximal SPTS, which correlated with the abnormal genoplot image identified. Interestingly, a proximal SPTS was confirmed in 19 of 20 SNPs (Tables 1 and 2). Of these 19 proximal SPTS, 9 proximal SPTS are novel SNPs and are not deposited in the dbSNP. Small deletion (7 bases) present in the probe-target sequence was confirmed in one SNP, rs1014824 (Table 1).

Table 1. Confirmation of a proximal probe-target sequence (SPTS) in 10 single nucleotide polymorphisms (SNPs) showing abnormal genoplot image.

Target SNP	Chr	SPTS	Relative distances (bp)	
			dbSNP[a]	Resequencing[b]
rs6431746	2	Novel	23	3 and 23
rs7677996	4	Novel	37	4 and 37
rs2245050	5	Novel	17	1 and 17
rs4896566	6	Novel	25	6 and 25
rs2543046	8	Novel	48	4 and 48
rs2441706	8	Novel	15	1 and 15
rs1962249	10	Novel	32	3 and 32
rs3818246	14	Novel	49	2 and 49
rs4786015	16	Novel	49	1 and 49
rs1014824	18		45	7-bp deletion

[a]Relative distances between targeted SNP and an SPTS deposited in NCBI dbSNP. [b]Relative distances between target SNP and an SPTS that was confirmed by direct sequencing.

Table 2. Effects of probe-target sequence (SPTS) on genotype detections in 10 single nucleotide polymorphisms (SNPs) showing an abnormal genoplot image.

Target SNP	Chr	SPTS	Relative distance	D’	n	Genotype combination between target SNP and SPTS {Error (%) /Missing (%)}		
						Ht *vs* P/M[a]	Ho *vs* M/M[b]	Other genotype combinations (Ho or Ht/P/P, Ho/Ht)
rs7424350	2	rs34721305	7	1	97	100/0	0/100	0/0
rs4373124	4	rs12645377	1	1	97	100/0	0/100	0/0
rs1325065	6	rs9404149	2	1	97	100/0	0/100	0/0
rs6954269	7	rs6953985	1	1	97	100/0	0/100	0/0
rs2543046	8	rs35615372	4	1	97	100/0	0/100	0/0
rs7079697	10	rs11200310	1	1	97	100/0	0/100	0/0
rs7103853	11	rs4756745	2	1	97	100/0	0/100	0/0
rs9538229	13	rs7326315	1	1	97	100/0	0/100	0/0
rs845567	20	rs6109557	2	1	97	100/0	0/100	0/0
rs396999	21	rs35907272	2	1	97	100/0	0/100	0/0

Ht and Ho = heterozygote and homozygote of target SNP, respectively. P/M, M/M and P/P = mismatch allelic heterozygote, mismatch allelic homozygote and perfect match allelic homozygote of SPTS, respectively. [a]Genotype combination of Ht *vs* P/M, and Ht are misclassified as Ho, at an error rate = 100%. [b]Genotype combination of Ho *vs* M/M, and Ho are called as missing, at a missing rate = 100%.

The Illumina Haman Hap300 platform uses relatively long probes (50 bases in length), which are less sensitive to single nucleotide mismatch, and hybridize immediately adjacent to the targeted SNP sites, followed by extension of a single nucleotide (Gunderson et al., 2005, 2006). Thus, a distal SPTS positioned at distances over 10 bases may allow efficient primer extension. However, a proximal SPTS is expected to affect hybridization affinities, leading to reduced allele-specific signal intensity measurements and abnormal genoplot image.

We noticed that the abnormal genoplot image is dependent on not only target SNP genotypes but also SPTS genotype (Figure 2A). The typical clusters were shown to correspond to an SPTS with perfect match allelic homozygote (P/P), while the extra-clusters were shown to correspond to an SPTS with mismatch allelic genotypes (P/M and M/M) (Figure 2A-C). The levels of signal intensity measurements (y-axis) somewhat reflected the genotypes, P/P, P/M and M/M, respectively (Figure 3A). Therefore, it is possible to deduce a set of genotypic imputation rules for SPTS through linkage disequilibrium pattern and cluster pattern (Franke et al., 2008). This direct detection of SPTS genotypes sometimes resulted in the identification of a novel SPTS, in perfect linkage disequilibrium with the target SNP (Table 1). This will be of particular interest in studies of exonic polymorphism, where rare SPTS could translate into phenotypic changes.

As with most genotyping platforms, it is difficult to distinguish heterozygous genotypes from homozygous genotypes in the case of a proximal SPTS (McCarroll et al., 2006). We found that the missing and erroneous genotypes occurred in fixed genotype combinations between target SNP and SPTS (Figure 2A-D). Following resequencing for 10 SNPs in 97 individuals, we found that individual genotypes with heterozygote (Ht) are misclassified as homozygote (Ho) in the presence of a proximal SPTS with heterozygote (P/M), at an error rate = 100%, while individual genotypes with homozygote (Ho) are called as missing in the presence of a proximal SPTS with mismatch allelic homozygote (M/M), at a missing rate = 100% (Table 2). No missing and erroneous genotypes were confirmed in the presence of a proximal SPTS with perfect match allelic homozygote (P/P) (data not shown).

In view of the perfect linkage disequilibrium (D' = 1) between proximal SPTS and target SNP (Table 2), it is possible to "correct" the genotypes of a target SNP from homozygote to heterozygote so that false-positive and false-negative associations can be prevented.

CONCLUSIONS

We demonstrated that proximal but not distal SPTS are associated with an abnormal genoplot image, which could result in spurious genotype detections. Undetected proximal SPTS are therefore likely to have an effect on the validity of SNP genotyping platforms, highlighting the importance of minimizing systematic errors in microarray experiments.

ACKNOWLEDGMENTS

We gratefully acknowledge the clinicians of Tokyo Women's Medical University who collected samples and obtained consent from sample donors. We express our appreciation to all study participants for donating their blood and time. We thank Dr. Christian Harkensee (University of Newcastle upon Tyne) for editing a number of drafts of this manuscript.

REFERENCES

Bemmo A, Benovoy D, Kwan T, Gaffney DJ, et al. (2008). Gene expression and isoform variation analysis using Affymetrix Exon Arrays. *BMC Genomics* 9: 529.

Benovoy D, Kwan T and Majewski J (2008). Effect of polymorphisms within probe-target sequences on oligonucleotide microarray experiments. *Nucleic Acids Res.* 36: 4417-4423.

Chen X and Kwok PY (1999). Homogeneous genotyping assays for single nucleotide polymorphisms with fluorescence resonance energy transfer detection. *Genet. Anal.* 14: 157-163.

Fan YS, Jayakar P, Zhu H, Barbouth D, et al. (2007). Detection of pathogenic gene copy number variations in patients with mental retardation by genomewide oligonucleotide array comparative genomic hybridization. *Hum. Mutat.* 28: 1124-1132.

Franke L, de Kovel CG, Aulchenko YS, Trynka G, et al. (2008). Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am. J. Hum. Genet.* 82: 1316-1333.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, et al. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37: 549-554.

Gunderson KL, Steemers FJ, Ren H, Ng P, et al. (2006). Whole-genome genotyping. *Methods Enzymol.* 410: 359-376.

Hwang DG and Green P (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101: 13994-14001.

Madsen BE, Villesen P and Wiuf C (2007). A periodic pattern of SNPs in the human genome. *Genome Res.* 17: 1414-1419.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, et al. (2006). Common deletion polymorphisms in the human genome. *Nat. Genet.* 38: 86-92.

Sapolsky RJ, Hsie L, Berno A, Ghandour G, et al. (1999). High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. *Genet. Anal.* 14: 187-192.

Sliwerska E, Meng F, Speed TP, Jones EG, et al. (2007). SNPs on chips: the hidden genetic code in expression arrays. *Biol. Psychiatry* 61: 13-16.

van Heel DA, Franke L, Hunt KA, Gwilliam R, et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* 39: 827-829.

Wang Y, Broderick P, Webb E, Wu X, et al. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* 40: 1407-1409.

Zhang L, Wu C, Carta R and Zhao H (2007). Free energy of DNA duplex formation on short oligonucleotide microarrays. *Nucleic Acids Res.* 35: e18.

Zhao Z and Boerwinkle E (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.* 12: 1679-1686.