ONLINE RESOURCES

Analysis of indel variations in the human disease-associated genes *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* in a Korean population

RYONG NAM KIM[1†], AERI KIM[1,2†], DONG-WOOK KIM[1†], SANG-HAENG CHOI[1], DAE-SOO KIM[1], SEONG-HYEUK NAM[1], ARAM KANG[1,2], MIN-YOUNG KIM[1], KUN-HYANG PARK[1], BYOUNG-HA YOON[1,2], KANG SEON LEE[1] and HONG-SEOG PARK[1,2*]

[1]*Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-806, Republic of Korea*

[2]*University of Science and Technology (UST), Daejeon 305-333, Republic of Korea*

[Kim R. N., Kim A., Kim D.-W., Choi S.-H., Kim D.-S., Nam S.-H., Kang A., Kim M.-Y., Park K.-H., Yoon B.-H., Lee K. S. and Park H.-S. 2012 Analysis of indel variations in the human disease-associated genes *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* in a Korean population. *J. Genet.* **91**, e1–e11. Online only: http://www.ias.ac.in/jgenet/OnlineResources/91/e1.pdf]

Introduction

Recently, the human genes *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* have emerged as important genetic factors that could be biologically associated with cancer, haematological diseases and olfactory dysfunction. In this regard, analysis of indel (insertion–deletion) variations in these genes at a population scale is of significant interest. In this study, we performed PCR amplifications and sequencing of the loci of these four genes using genomic DNA from 100 Korean individuals. We analysed the indels in these genes and made predictions about the functional consequences that are likely caused by the indels. We discovered a 3-bp deletion in *CDKN2AIP*, a 15-bp insertion in *WDR66*, a 3-bp deletion in *USP20* and a 3-bp insertion in *OR7C2* with indel allele frequencies of 33%, 79%, 99% and 100%, respectively. The results of 3-D structural predictions and analysis of the proteins encoded by the four genes showed that in-frame amino acid deletions or insertions caused by these indels could result in hindrance of the formation of the optimal functional structures of these proteins, which could affect their functions. In particular, the indels in *USP20* and *OR7C2* have severely biased frequencies in the investigated Korean population, which may reflect a high susceptibility to certain cancer types and a biased preference to a kind of smell in Korean individuals. Additionally, our results could help to identify therapeutic targets for treating possible genetic diseases in individuals possessing homozygous genotypes for these indels in future studies.

Next-generation sequencing has accelerated human genome sequencing with unprecedented speed at the population scale across diverse ethnic groups (Wang *et al.* 2008; Wheeler *et al.* 2008; Ahn *et al.* 2009; Kim *et al.* 2009). Analyses of single nucleotide polymorphisms (SNPs) in human populations have been reported relatively comprehensively in previous studies (The International HapMap Project 2003; McCarroll *et al.* 2008; Durbin *et al.* 2010), but investigations of indel polymorphisms (Mills *et al.* 2006; Bentley *et al.* 2008; Ye *et al.* 2009; Hajirasouliha *et al.* 2010; Mullaney *et al.* 2010) at the scale of human populations still lag far behind SNP analyses. At present, little is known about genomewide indel polymorphisms at the scale of the Korean population.

Previous studies have suggested that numerous indels are closely associated with human genetic diseases. One of the most common human genetic diseases, cystic fibrosis, is caused by an allele of the *CFTR* gene harbouring a three-base pair deletion that results in the elimination of a single amino acid from the encoded protein, which in turn leads to the disease (Collins *et al.* 1987). Fragile X syndrome had been found to be caused by DNA insertions derived from the expansion of short trinucleotide repeat units (Warren *et al.* 1987). DNA insertions produced by the integration of transposable genetic elements into the genome may cause human genetic diseases (Ostertag and Kazazian 2001). In particular, *Alu*, L1, and SVA transposon insertions have been reported to disrupt gene function and cause human diseases such as haemophilia, neurofibromatosis, muscular dystrophy and cancer (Ostertag and Kazazian 2001). Additionally, recent studies have reported that indel mutations are key drivers causing human cancers (Ngo *et al.* 2011; Varela *et al.* 2011). Such prevalent and close associations between indel

*For correspondence. E-mail: hspark@kribb.re.kr.
†These authors contributed equally to this work.

Keywords. insertion–deletion variations; haematological disease; tumours; human genetics.

polymorphisms and human genetic diseases suggest that there is an urgent need to speed up comprehensive indel analyses, not only across diverse ethnic groups, but also at the scale of whole populations.

In the present study, we carried out analyses to obtain putative functional insights related to indel variations present in the human genes *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* at the scale of a Korean population. A previous study by our group (Kim *et al.* 2010) aimed at capturing and analysing the whole exome in a Korean genome using array-based hybridization techniques and next-generation sequencing technology, and we observed that the above-mentioned four genes, which have recently been reported to be biologically associated with causes of human diseases (Malnic *et al.* 2004; Li *et al.* 2005; Berthouze *et al.* 2009; Cheung *et al.* 2009; Meisinger *et al.* 2009; Soranzo *et al.* 2009), harboured indel variations that appeared to be nearly exclusive to the Korean genome based on an *in silico* comparison with representative genomes of other ethnic groups. The present study shows, for the first time, the allele and genotype frequencies of the genes harbouring these indels in a Korean population and makes predictions regarding their functional consequences.

Materials and methods

Genomic DNA

We isolated genomic DNA from blood samples of 100 randomly selected healthy Korean individuals using a blood genomic DNA extraction kit (QG-Mini80, Fujifilm, Tokyo, Japan).

PCR amplification and cloning

To amplify the local genomic regions harbouring indels within the loci of the four genes in the 100 Korean genomes, we designed PCR primers, as shown in table 1. PCR amplifications were performed in a 20-μL reaction volume containing 1 μL of gDNA, 0.5 μL of *Taq* DNA polymerase, 2 μL of 10× *Taq* buffer, 0.5 μL of 2.5 mM dNTP mix, 1 × 2 μL of both 3.5 pmol primers (forward and reverse) and 14 μL of distilled water. The reactions were performed under the following conditions: an initial hot start incubation at 95°C for 5 min; 35 cycles of a denaturing step at 95°C (20 s), an annealing step at 62°C (20 s) and an elongation step at 72°C (30 s); and a terminal incubation step at 72°C for 7 min. The resulting PCR products were ligated into a T and A cloning vector (Real Biotech Corporation, Taipei, Taiwan) by incubating at 4°C overnight in a 10-μL reaction volume. The *Escherichia coli* DH5α strain was transformed with 5 μL of the ligation product. Each insert was sequenced using one M13 forward primer and one M13 reverse primer complementary to sequences present in the RBC T and A cloning vector.

Table 1. Primer list.

Primer name	Primer sequence
CDKN1AIP-F	GAGCTCTGGCATCTCCAGTC
CDKN1AIP-R	ACAAGGGCACCTCGATCTCT
WDR66-F	CCGAGAAGCAACAGGAGAAA
WDR66-R	CTGTGTCTCCAAACGGATCA
USP20-F	CTGAGAAGGAGCGGATGAAG
USP20-R	CGGTAGGTACTGGGCCATAG
OR7C2-F	GTGTTTTTCCTCTCTGTGGA
OR7C2-R	AGATAGACCCCAAGGCCAGT

Additionally, to identify the heterozygous genotype of each indel at the sequence level, we directly sequenced the genomic DNA PCR products harbouring the indel regions without any further cloning steps. Thus, if a Korean individual was heterozygous for an indel, the genomic DNA PCR products would simultaneously include deletion and insertion allele types of the indel, resulting in mixed chromatogram signals corresponding to sequences of both alleles (figure 1B).

Sanger capillary sequencing

Using the prepared plasmid DNAs harbouring inserts as templates, sequencing reactions were performed according to the manufacturer's instructions using BigDye Terminator chemistry v3.1; Cycle Sequencing kit (Applied Biosystems, Foster City, USA). PCR amplification was performed in 3-μL volumes containing 250 ng of plasmid DNA, 0.5 μL of a universal primer (3 pmol), 0.6 μL of 5× sequencing buffer, 1.65 μL of distilled water and 0.25 μL of BDT v3.1, using a GeneAmp PCR System 2720 (Applied Biosystems, Foster City, USA). The reactions were performed under the following conditions: 35 cycles of denaturation at 96°C for 10 s, annealing at 50°C for 5 s, and extension at 60°C for 4 min. PCR products were purified via ethanol precipitation and resolved on an ABI 3730XL DNA analyzer (Applied Biosystems, Foster City, USA). Sequencing data were assembled and analysed using the Sequencher 4.1.5 program (Gene Codes, Ann Arbor, USA).

Bioinformatics annotation and 3-D protein structure analysis

In this study, the UCSC genome browser (http://genome.ucsc.edu/), UCSC proteome browser (http://genome.ucsc.edu/cgi-bin/pbGateway), InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/), Prosite database (http://expasy.org/prosite/) and Phyre program (Kelley and Sternberg 2009) (http://www.sbg.bio.ic.ac.uk/phyre/) were used to display the gene structures at the genomic loci and to predict protein domain structures and 3-D protein structures.

Gel electrophoresis to identify indel genotypes

To visibly separate the PCR-amplified genomic DNA fragments obtained using genomic DNA from blood samples

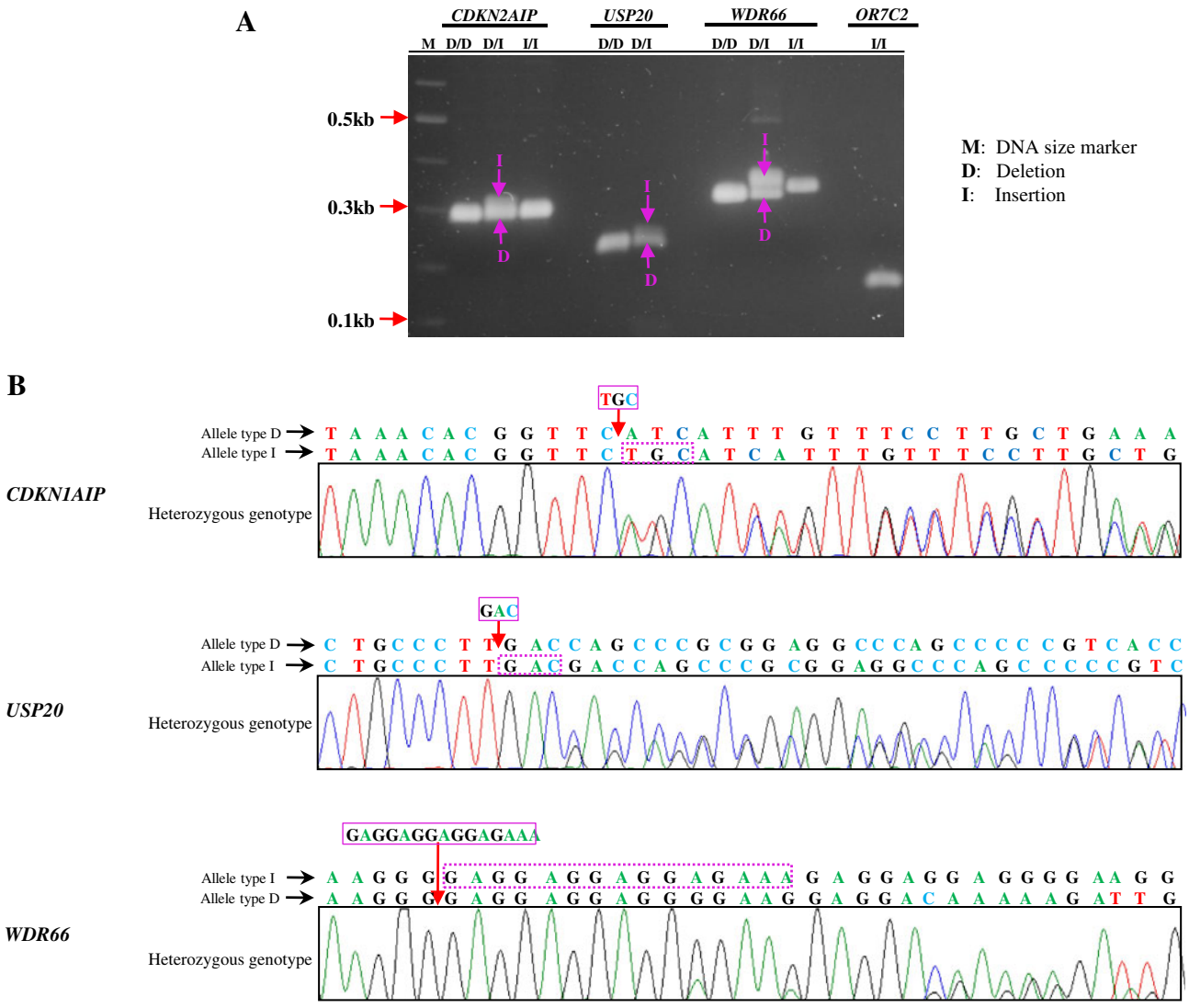

Figure 1. PCR-amplifications of local genomic regions harbouring the indels. **A**) Following electrophoresis in 2% agarose gel for 3 h, bands that were different from each other by only 3-bp or 15-bp in length were visibly separated on the gel. Pink arrowheads indicate bands (I) with an insertion and bands (D) with a deletion, which correspond to each local genomic region. **B**) Heterozygous genotypes of the indels in *CDKN2AIP*, *USP20* and *WDR66*. The regional DNA sequences surrounding and corresponding to a deletion (D)/insertion (I) are shown. Sequence chromatograms corresponding to genomic DNA PCR products (including deletion and insertion allele types) obtained from the genomes of Korean individuals heterozygous for the indels are also shown.

of Korean individuals heterozygous for a given insertion or deletion, gel electrophoresis was performed on 2% agarose gel for 3 h.

Results

Allele and genotype frequencies

By PCR-amplifying and sequencing sub-local genomic regions harbouring indels in the genomic DNA of 100 Korean individuals obtained using blood samples, we determined the allele and genotype frequencies of indels at the genomic loci of *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* (table 2; figures 1, A&B). The *CDKN2AIP* locus was found to harbour a 3-bp deletion (c.723_725delTGC) within exon 3, which was present in the proportions of 0.67 (+) and 0.33 (−) with respect to allele frequency and 0.44 (+/+), 0.47 (+/−) and 0.09 (−/−) with respect to genotype frequency in the Korean population. A 15-bp insertion (c.186_187insGAGGAGGAGGAGAAA) in exon 2 of the *WDR66* gene was present at allele frequencies of 0.79 (+) and 0.21 (−) and genotype frequencies of 0.63 (+/+), 0.32 (+/−) and 0.05 (−/−). In contrast to the very low frequencies of homozygous deletion genotypes found for the two above-mentioned indels, a 3-bp deletion (c.1075_1077delGAC) in exon 11 of the *USP20* gene

Table 2. Allele and genotype frequencies of indels.

Gene	Type of indels		Indel length	Location	Allele frequencies		Genotype frequencies			Tested number
	cDNA level	Protein level			+	−	++	+−	−	
CDKN2AIP	c.723_725delTGC	p.A242del	3 bp[a]	exon 3	0.67	0.33	0.44	0.47	0.09	96
WDR66	c.186_187insGAGGAGGAGGAGAAA	p.G62_E63insEEEEK	15 bp[b]	exon 2	0.79	0.21	0.63	0.32	0.05	89
USP20	c.1075_1077delGAC	p.D359del	3 bp[c]	exon 26	0.01	0.99		0.02	0.98	98
OR7C2	c.683_684insATC	p.V228_S229insS	3 bp[d]	exon 1	1		1			98

a, TGC; b, GAGGAGGAGGAGAAA; c, GAC; d, ATC; +, present; −, absent.

was present at a high frequency in the investigated Korean population: allele frequencies of 0.01 (+) and 0.99 (−) and genotype frequencies of 0.02 (+/−) and 0.98 (−/−). The +/+ genotype of this 3-bp deletion appeared to be nearly absent from this Korean population (figure 1A; table 2). Additionally, the *OR7C2* gene locus was found to harbour a 3-bp insertion (c.683_684insATC) that appeared to be present in homozygosity in the whole Korean population (1 (+) for allele frequency and 1 (+/+) for genotype frequency). These results suggest that the allele and genotype frequencies of the indels present within the four investigated genes are highly diverse. Next, we focussed on obtaining detailed insights into the functional consequences likely caused by the indel variations in these genes.

In-frame deletion in CDKN2AIP

In recent years, the gene *CDKN2AIP* (also called *CARF* or *FLJ20036*) has been reported to be an emerging regulator of the p53 tumour suppressor and senescence pathway (Cheung *et al.* 2009). Thus, the indel variation present in this gene is noteworthy because of the important biomedical roles of the human protein CARF.

The human *CDKN2AIP* gene is located at q35.1 locus on chromosome 4. We found that the 3-bp deletion (c.723_725delTGC) observed within exon 3 of this gene resulted in an in-frame deletion (p.A242del) of A242, which is embedded in a domain (129–578) that is similar to NF-kappaB repressing factor (figure 2A, B&C). The deleted 3-bp nucleotide sequence TGC (c.723_725delTGC) in the *CDKN2AIP* gene is composed of 723th single nucleotide T (that is participated in encoding an amino acid S241) and two nucleotides G (724th) and C (725th) that are participated in encoding an amino acid A242. In the case of *CDKN2AIP* gene having the deletion (c.723_725delTGC), the codon sequence TCT encoding the S241 is changed into TCA by integrating 726th nucleotide A that was participated in encoding the A242. However, because the codon sequences TCT and TCA, both encode the amino acid residue serine (S), finally the deletion (c.723_725delTGC) results in only the deletion (p.A242del) of A242 in protein level but 241th amino acid remains still S241. The domain containing A242 is adjacent to a domain of unknown function, DUF3469 (19–125). In addition to these domains, the CDKN2AIP protein encoded by this gene harbours the domains DS-RNA-bd (461–536) (double-stranded RNA-binding) and NLS_BP (537–551) (bipartite nuclear localization signal).

Notably, A242 has been evolutionarily well conserved among humans, primates and mammals, but is not conserved in other species (figure 2D), implying that this amino acid could be essential in forming a functionally optimal structural conformation of this protein, which might exhibit biological functions that are more important for evolutionarily higher eukaryotes.

Our analysis of the predicted 3D structure of the amino acid sequence of this protein from Q172 to A448 using the

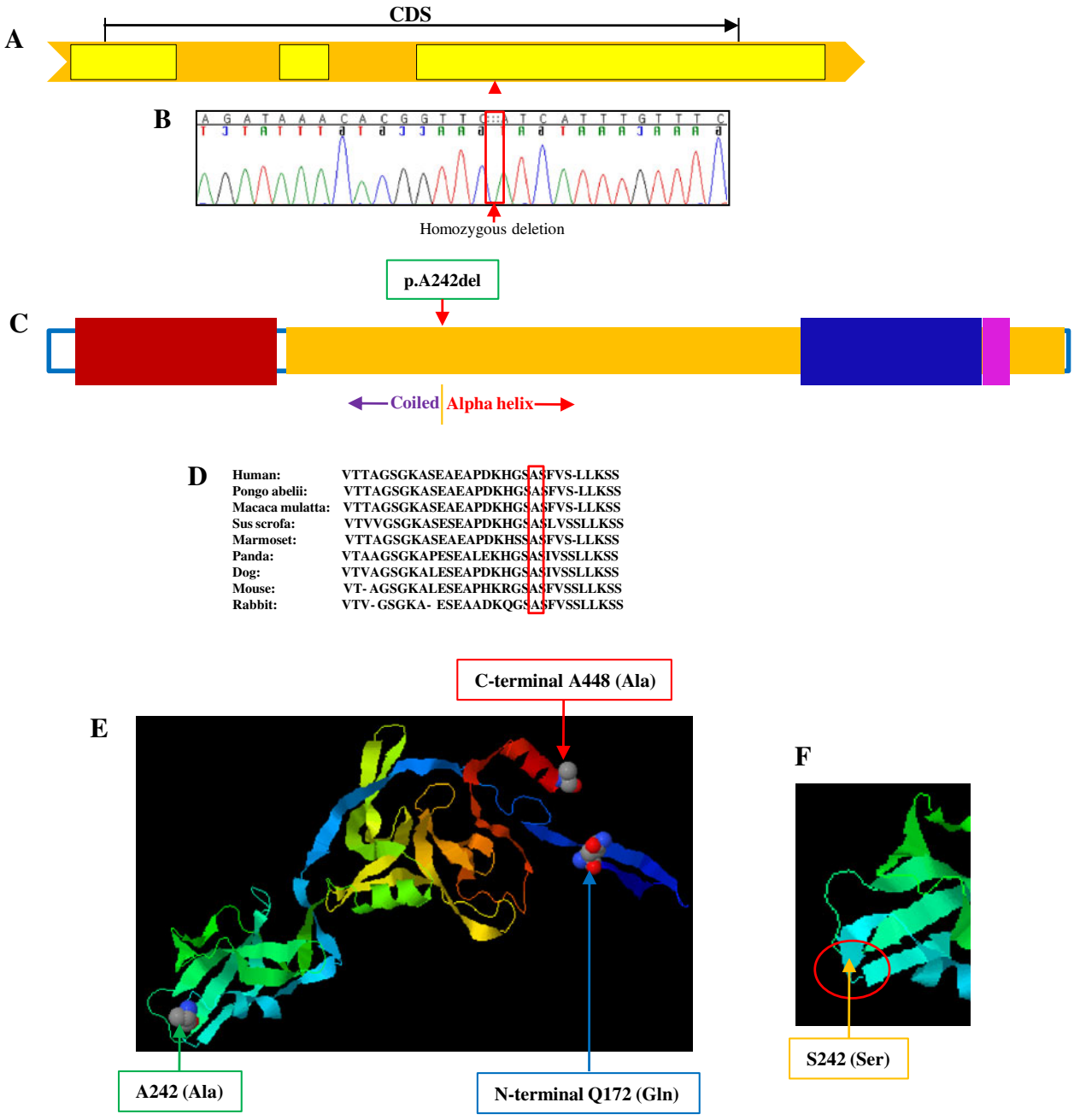

Figure 2. In-frame deletion of A242 in CDKN2AIP. A) Genomic structural organization of the *CDKN2AIP* gene. The yellow boxes indicate exons. The orange arrow bar covers the genomic locus of the *CDKN2AIP* gene, including intron regions, indicating the direction of transcription. The red triangle indicates the position of the 3-bp deletion. B) Sequence chromatogram: the boundary of red rectangle indicates the position of the 3-bp deletion. C) Protein domain structure of CDKN2AIP: the boundary of the light blue rectangle represents the CDKN2AIP protein. The red and yellow boxes indicate the domain DUF3469 (19–125), whose function is unknown, and a domain (129–578) similar to NF-kappaB repressing factor, respectively. The dense blue and pink boxes, which are embedded in the yellow domain box, indicate a double-stranded RNA-binding domain (DS-RNA-bd) (461–536) and a bipartite nuclear localization signal (NLS_BP) (537–551), respectively. D) Evolutionary conservation of the alanine residue A242 among human, primate and mammalian species. The boundary of the red rectangle indicates high conservation of the A242 among these species. E) Three-dimensional structure corresponding to the amino acid sequence from Q172 to A448 of the normal CDKN2AIP protein including A242. F) Subregion of the 3-D structure of the CDKN2AIP protein after deletion of A242. An enlargement of the micro-3-D structural region harbouring S242, which replaces A242, is shown within a red circular line.

Phyre program (http://www.sbg.bio.ic.ac.uk/phyre) showed that the deletion of A242 could cause a decrease in a micro-helical region at the joint border between the helix and coiled regions (figure 1, E&F), which might directly or indirectly affect the 3-D structural conformation of this protein. Although previous studies have reported that the

CARF protein has several functional domains that are helpful in binding not only to ARF but also to p53 to activate the p53 pathway via ARF activation and HDM2 inactivation (Cheung *et al.* 2009), our domain annotation showed that this protein contains a large domain (129–578) of as-yet unknown function similar to the NF-kappaB repressing factor domain (figure 2C). If NF-kB inactivation by CARF is demonstrated in the future, the mechanism of the arrest of cellular proliferation caused by CARF could be further elucidated through new molecular insights related to an interaction between NF-kB and CARF, which has not been found to date. In line with such functional considerations, the deletion of A242, which is embedded in the NF-kB repressing factor-like domain, could affect protein–protein interactions of CARF with other proteins that are crucial in maintaining normal cellular function.

In-frame insertion in WDR66

Recent genomewide association studies have revealed that *WDR66* (also called *MGC33630* or *FLJ39783*) is very closely associated with mean platelet volume (MPV), which is increased during myocardial and cerebral infarction and is a strong, independent predictor of postevent morbidity and mortality (Meisinger *et al.* 2009; Soranzo *et al.* 2009). Therefore, analysis of indel variations in the human *WDR66* gene, which resides at the locus q24.31 on chromosome 12, could be important in understanding haematological genetic diseases.

The human *WDR66* is transcribed into two variants: variant 1 (1NM_144668) and variant 2 (NM_001178003). The 15-bp insertion (c.186_187insGAGGAGGAGGAGAAA) identified in exon 2 of *WDR66* in this study results in an in-frame insertion (p.G62_E63insEEEEK) of five amino acids in two protein isoforms corresponding to transcript variants 1 and 2 of this gene (figure 3, A, B&C).

Using the InterPro domain annotation program (http://www.ebi.ac.uk/interpro/), we found that nine WD40 repeat domains are present in both isoforms 1 and 2, and one EF-hand domain is present only in isoform 2. To gain a clear insights related to the 3-D structure of this protein, we obtained a predicted 3-D structure of the amino acid sequence (from W257 to T922) corresponding to WDR66 protein isoforms 1 and 2 using the Phyre program (figure 3D). The predicted 3D structure of the peptide fragment from the WDR66 protein presented a circularized propeller form consisting of six β-sheet blades, which is very typical of WD40 repeat-containing proteins (Li and Roberts 2001). Unfortunately, the in-frame insertion (p.G62_E63insEEEEK) region could not be included when drawing the predicted 3-D structure shown in figure 3D due to deficiency of other representative 3-D models corresponding to the insertion-containing region of the WDR66 protein sequence. Regardless, this insertion, which resides outside the nine WD40 repeat domains in WDR66, could have a large effect on the biological functions of this protein (Li and Roberts 2001). Researchers previously reported finding that WD40 repeat domain-containing proteins are involved in many essential biological functions, ranging from signal transduction and transcription regulation to apoptosis, and these domains have also been associated with several human diseases (Li and Roberts 2001). Importantly, the results of previous studies have suggested that the functional specificities of these proteins are determined by sequences outside the WD40 repeats. Nevertheless, several interesting questions related to how the 3-D conformational change caused by the insertion could affect the specificity of WDR66 in mutually or multi-laterally interacting with other proteins remain to be answered in future studies.

In-frame deletion in USP20

The deubiquitinating enzyme, USP20, was first discovered as a substrate of the von Hippel–Lindau tumour suppressor protein (pVHL), and mutation of the *pVHL* gene has been found to be strongly associated with a variety of tumours, including haemangioblastomas in the central nervous system and retina, clear cell carcinomas of the kidney, pheochromocytomas of the adrenal gland, and cysts/adenomas and islet cell tumours of the pancreas (Goldsmith and Thomas 1999; Li *et al.* 2002). Recently, it was reported that by ubiquitinating USP20 and inducing its degradation, pVHL can inhibit the activity of HIF-1α, which can be deubiquitinated and stabilized by USP20 (Li *et al.* 2005). Additionally, USP20 can serve as a novel regulator that dictates both postendocytic sorting and the intensity and extent of b2AR signalling from the cell surface (Berthouze *et al.* 2009).

The human *USP20* (also called *VDU2*, *hVDU2*, *KIAA1003* or *LSFR3A*) gene resides at the q34.11 locus on chromosome 9. This gene is transcribed into three variants: variant 1 (NM_006676), variant 2 (NM_001008563) and variant 3 (NM_001110303). The 3-bp deletion (c.1075_1077delGAC) identified in this study is located in exon 11 of the three transcript variants originating from this gene, causing an in-frame deletion (p.D359del) at the amino acid sequence level (figure 4, A, B, C, D&E). We found that the USP20 protein has three functional domains: a zinc finger (Znf) domain (30–91), a peptidase C19 domain (143–682) and an ubiquitin-specific peptidase (DUSP) domain (702–785 and 810–895). Of particular interest, D359 is located between two important functional motifs, UCH_2_1 (ubiquitin carboxyl-terminal hydrolases family 2 signature 1) and UCH_2_2 (ubiquitin carboxyl-terminal hydrolases family 2 signature 2), which have been shown to play crucial roles in the deubiquitinating activity of USP20 (figure 4E). Our analysis of the predicted 3-D structure of the amino acid sequence between P141 and S686 in the USP20 protein showed that the UCH_2_1 (from G146 to Q161) and UCH_2_2 (from Y627 to Y644) regions exhibited helix-turn-helix and β-sheet structures, respectively (figure 4F).

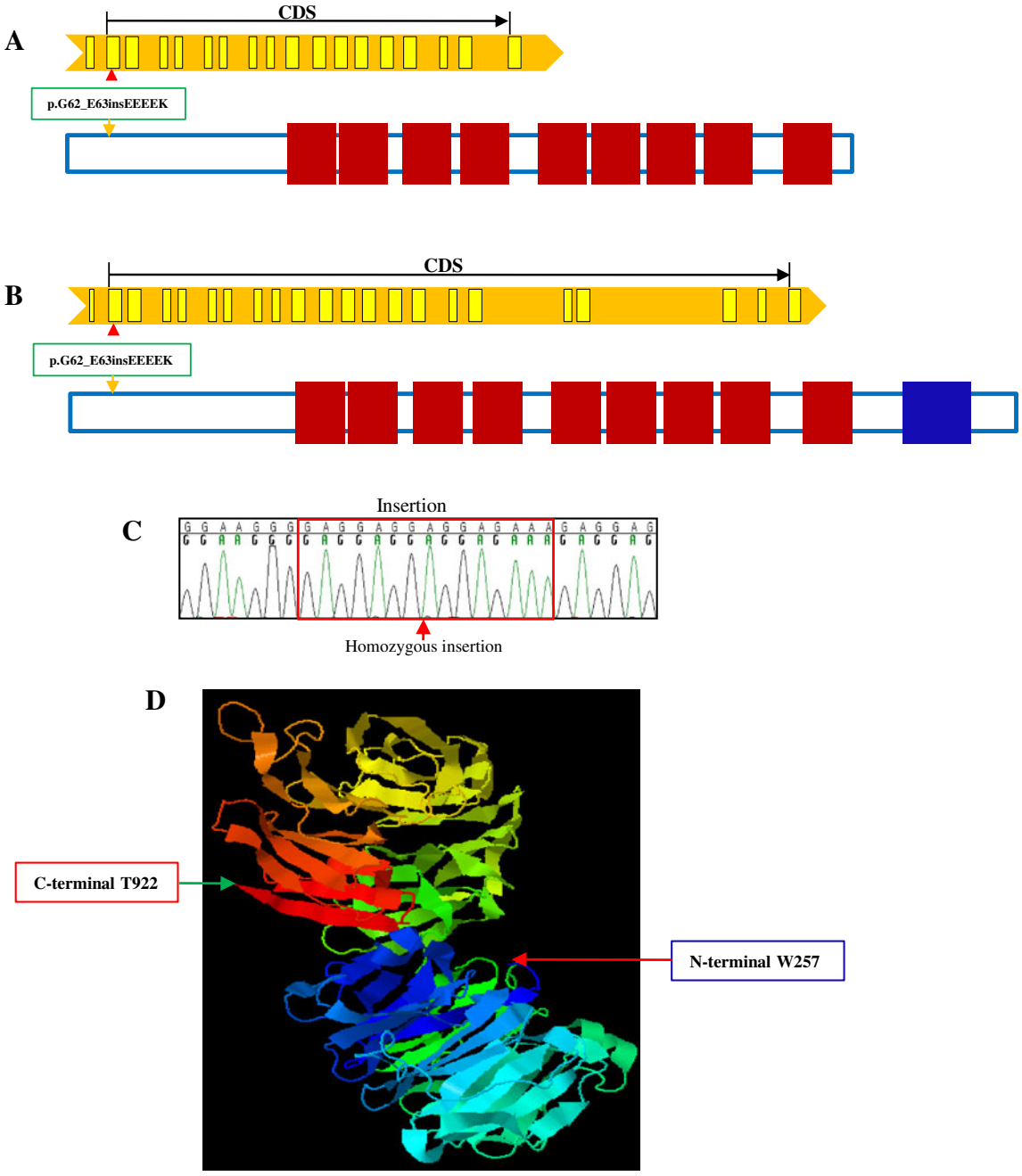

Figure 3. In-frame insertion in WDR66. A) Genomic structural organization and protein domain structure corresponding to transcript variant 2 (NM_001178003) of the human *WDR66* gene. The nine red boxes indicate WD40 repeat domains (288–329, 332–376, 435–474, 493–527, 595–634, 639–677, 682–721, 730–773 and 880–920). B) Genomic structural organization and protein domain structure corresponding to transcript variant 1 (1NM_144668) of *WDR66*. Nine red boxes indicate the WD40 repeat domains, which are located at the same locations as those of isoform 2. The dense blue box indicates an EF-hand domain (980–1036). The WDR66 protein isoforms 1 and 2 include the same insertion (p.G62_E63insEEEEK) at the same position. The red triangles indicate the positions of the 15-bp insertion in transcript variants 1 and 2. C) Sequence chromatogram showing the inserted nucleotide bases within the boundary of the red rectangle. D) Three-dimensional structure corresponding to the amino acid sequence from W257 to T922 of WDR66 protein isoform 1.

Unfortunately, we could not accurately predict a 3-D structure of the deletion region lacking D359 because of a lack of other representative protein models corresponding to this region in the USP20 protein. Interesting questions related to how the D359 deletion could directly or indirectly affect the formation of the optimal structural conformations of the two signatures UCH_2_1 and UCH_2_2 for the deubiquitination reaction remain to be addressed in future studies.

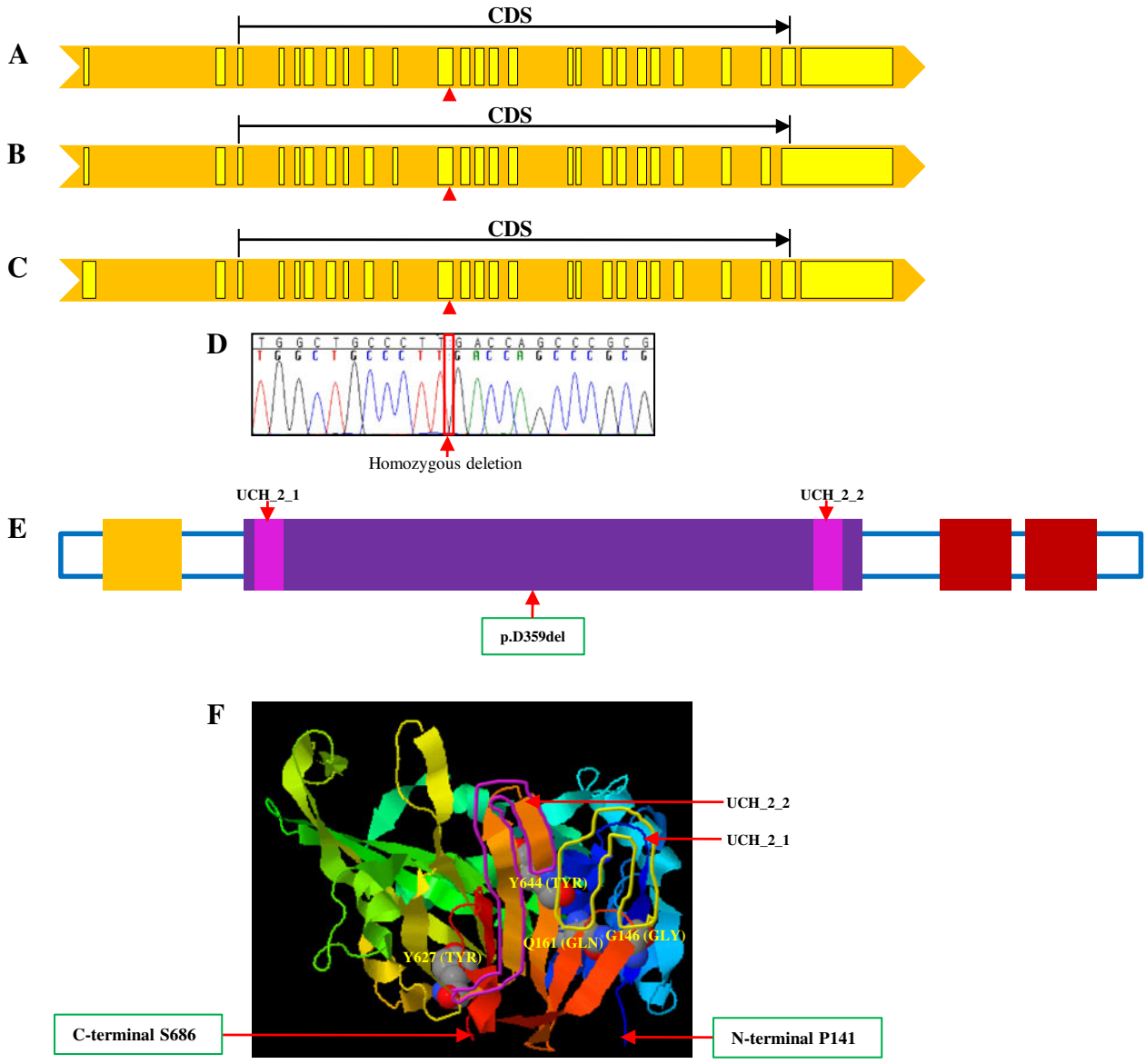

Figure 4. In-frame deletion in USP20. A), B) and C) show the genomic structural organizations corresponding to transcript variants 3 (NM_001110303), 1 (NM_006676), and 2 (NM_001008563) of the *USP20* gene, respectively. The red triangles indicate the positions of the same deletion in each of the three transcript variants. D) Sequence chromatogram showing the 3-bp deletion. E) Domain structure of the USP20 protein. The orange, purple and red boxes indicate a zinc finger (Znf) domain (30–91), a peptidase C19 domain (143–682) and a ubiquitin-specific peptidase (DUSP domain) (702–785 and 810–895), respectively. The two pink boxes embedded in the purple box indicate UCH_2_1 (ubiquitin carboxyl-terminal hydrolases family 2 signature 1) and UCH_2_2 (ubiquitin carboxyl-terminal hydrolases family 2 signature 2). F) Three-dimensional structure corresponding to the amino acid sequence between P141 and S686 of the USP20 protein. Two sub-regional 3-D structures corresponding to the UCH_2_1 (between G146 and Q161) and UCH_2_2 (between Y627 and Y644) signatures are shown within yellow and pink lines, respectively.

In-frame insertion in OR7C2

Olfactory receptors interact with odorant molecules in the nose to initiate a neuronal response that triggers the perception of a smell. In general, olfactory receptor proteins are members of a large family of G-protein-coupled receptors (GPCR) arising from single exon genes (Malnic *et al.* 2004).

The olfactory receptor gene *OR7C2* (also called *OR7C3*, *OR19-18* or *CIT-HSP-87M17*) resides at the p13.12 locus on human chromosome 19. The 3-bp insertion (c.683_684insATC) in *OR7C2* identified in this study results in an in-frame insertion (p.V228_S229insS) of S229, which is located between the fourth rhodopsin-like GPCR superfamily domain and the fourth olfactory receptor domain in the protein encoded by this gene (figure 5, A, B&C). Notably, the inserted S229 between the fifth and sixth transmembrane domains appears to reside at the inner side of the membrane, facing the cytoplasm (figure 5D).

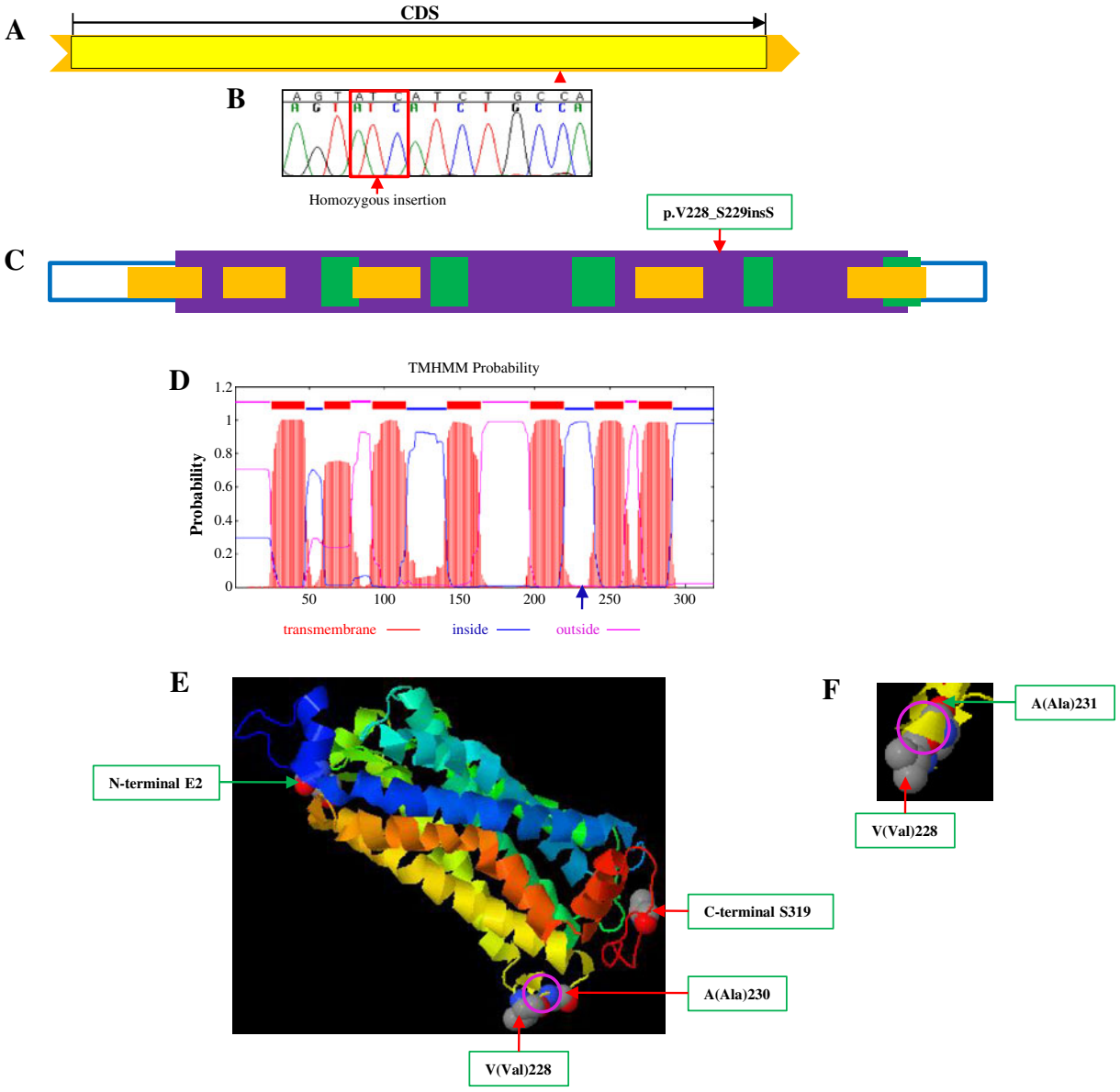

Figure 5. In-frame insertion in OR7C2. A) Genomic structural organization of the human *OR7C2* gene. B) Sequence chromatogram showing the 3-bp insertion. C) Domain structure of the OR7C2 protein. The purple box indicates the 7TM_GPCR_Rhodpsn domain (41–289). The green and orange boxes embedded in the purple box, some of which overlap, indicate olfactory receptor domains (92–103, 129–141, 176–192, 235–244 and 282–293) and rhodopsin-like GPCR superfamily domains (26–50, 59–80, 104–126, 199–222 and 271–297), respectively. D) TMHMM probability of seven transmembrane domain structures. The blue arrowhead indicates the position of the S229 insertion, which resides on the intracellular side of the protein. E) Three-dimensional structure corresponding to the amino acid sequence between E2 and S319 of the OR7C2 protein. The micro-3-D structural region between V228 and A230 without the S229 insertion is shown within a pink circular boundary line. F) The micro-3-D structural region surrounding the inserted S229 residue (within a pink circular boundary line) between V228 and A231.

To visualize the conformational structural change caused by the S229 insertion, we depicted the predicted 3-D structure of OR7C2 corresponding to the amino acid sequence between E(Glu)2 and S(Ser)319 (figure 5, E&F). From this analysis, we observed that a coiled structure (within the pink circular boundary in figure 5E) present between V(Val)228 and A(Ala)230 when there was not insertion of S229 was altered to a helical structure (within the pink circular boundary in figure 5F) between V(Val)228 and A(Ala)231 following this insertion. However, whether the conversion from a coiled structure to a helix due to the S229 insertion into this microregion would seriously affect the molecular function of the OR7C2 protein remains uncertain at the present time. Nevertheless, given that OR7C2 is an olfactory receptor, the slight bias in the interaction with odorant molecules caused by this very slight conformational

structural change could result in signals being skewed significantly from the norm during neuronal signal transmission related to the perception of a smell.

Discussion

In this study, we identified the allele and genotype frequencies of indels in the human genetic disease-associated genes *CDKN2AIP*, *WDR66*, *USP20* and *OR7C2* at the scale of a Korean population for the first time and predicted the functional consequences of these variations.

Among the four indels we identified, the 3-bp deletion (c.1075_1077delGAC) in *USP20* and the 3-bp insertion (c.683_684insATC) in *OR7C2* appear to be nearly exclusively present in the Korean population. Although we found that the four tested indels were biased towards Korean genomes (Kim *et al.* 2009, 2010; Ahn *et al.* 2009) compared with publically available representative genome data from individuals of other ethnic backgrounds, no previous reports exist about the allele and genotype frequencies of these INDELs at the scale of other ethnic population.

Another important finding of this study was the high frequency of the 3-bp deletion in the *USP20* gene in Korean individuals (allele frequency, 0.99; homozygous genotype frequency, 0.98). Given that a functional deficiency of USP20 could result in carcinogenesis due to improper mediation of the tumour suppression mechanism related to pVHL (Goldsmith and Thomas 1999; Li *et al.* 2002), this high frequency of 3-bp deletion in the Korean population could imply a biased susceptibility of Korean individuals to certain cancer types compared with other ethnic groups. However, this inference remains to be proven, as there are still no available data on the allele and genotype frequencies of the 3-bp deletion in the gene *USP20* at the scale of other ethnic population.

Nevertheless, the allele and genotype frequencies of the indels described here and the related predicted functional consequences elucidated in this study present new opportunities for future studies on genetic diseases that are likely to be more prevalent in Korean populations than in other ethnic groups and for the search for drug targets for the treatment of these diseases.

Acknowledgements

We thank the members of the Genome Resource Center (GRC) in the Korea Research Institute of Bioscience and Biotechnology for their active assistance in performing this research. This research was supported by grant 2009-0084206 from the Ministry of Education, Science and Technology and grant KGM5411011 from KRIBB.

References

Ahn S. M., Kim T. H., Lee S., Kim D., Ghang H., Kim D. S. *et al.* 2009 The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629.

Bentley D. R., Balasubramanian S., Swerdlow H. P., Smith G. P., Milton J., Brown C. G. *et al.* 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59.

Berthouze M., Venkataramanan V., Li Y. and Shenoy S. K. 2009 The deubiquitinases USP33 and USP20 coordinate beta2 adrenergic receptor recycling and resensitization. *EMBO J.* **28**, 1684–1696.

Cheung C. T., Hasan M. K., Widodo N., Kaul S. C. and Wadhwa R. 2009 CARF: an emerging regulator of p53 tumor suppressor and senescence pathway. *Mech. Ageing Dev.* **130**, 18–23.

Collins F. S., Drumm M. L., Cole J. L., Lockwood W. K., Vande Woude G. F. and Iannuzzi M. C. 1987 Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**, 1046–1049.

Durbin R. M., Abecasis G. R., Altshuler D. L., Auton A., Brooks L. D., Gibbs R. A. *et al.* 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.

Goldsmith D. J. and Thomas P. 1999 Images in clinical medicine. von Hippel-Lindau disease. *N. Engl. J. Med.* **340**, 1880.

Hajirasouliha I., Hormozdiari F., Alkan C., Kidd J. M., Birol I., Eichler E. E. and Sahinalp S. C. 2010 Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* **26**, 1277–1283.

Kelley L. A. and Sternberg M. J. 2009 Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371.

Kim D. W., Nam S. H., Kim R. N., Choi S. H. and Park H. S. 2010 Whole human exome capture for high-throughput sequencing. *Genome* **53**, 568–574.

Kim J. I., Ju Y. S., Park H., Kim S., Lee S., Yi J. H. *et al.* 2009 A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015.

Li D. and Roberts R. 2001 WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell Mol. Life Sci.* **58**, 2085–2097.

Li Z., Wang D., Na X., Schoen S. R., Messing E. M. and Wu G. 2002 Identification of a deubiquitinating enzyme subfamily as substrates of the von Hippel-Lindau tumor suppressor. *Biochem. Biophys. Res. Commun.* **294**, 700–709.

Li Z., Wang D., Messing E. M. and Wu G. 2005 VHL protein-interacting deubiquitinating enzyme 2 deubiquitinates and stabilizes HIF-1alpha. *EMBO Rep.* **6**, 373–378.

Malnic B., Godfrey P. A. and Buck L. B. 2004 The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. USA* **101**, 2584–2589.

McCarroll S. A., Kuruvilla F. G., Korn J. M., Cawley S., Nemesh J. and Wysoker A. 2008 Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40**, 1166–1174.

Meisinger C., Prokisch H., Gieger C., Soranzo N., Mehta D., Rosskopf D. *et al.* 2009 A genome-wide association study identifies three loci associated with mean platelet volume. *Am. J. Hum. Genet.* **84**, 66–71.

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.

Mullaney J. M., Mills R. E., Pittard W. S. and Devine S. E. 2010 Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19**, R131–R136.

Ngo V. N., Young R. M., Schmitz R., Jhavar S., Xiao W., Lim K. H. *et al.* 2011 Oncogenically active MYD88 mutations in human lymphoma. *Nature* **470**, 115–119.

Ostertag E. M. and Kazazian Jr H. H. 2001 Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538.

Soranzo N., Spector T. D., Mangino M., Kuhnel B., Rendon A., Teumer A. *et al.* 2009 A genome-wide meta-analysis identifies

22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41**, 1182–1190.

The International HapMap Project 2003 *Nature* **426**, 789–796.

Varela I., Tarpey P., Raine K., Huang D., Ong C. K., Stephens P. *et al.* 2011 Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–542.

Wang J., Wang W., Li R., Li Y., Tian G., Goodman L. *et al.* 2008 The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65.

Warren S. T., Zhang F., Licameli G. R. and Peters J. F. 1987 The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science* **237**, 420–423.

Wheeler D. A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A. *et al.* 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876.

Ye K., Schulz M. H., Long Q., Apweiler R. and Ning Z. 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871.

Received 14 June 2011, in final revised form 26 September 2011; accepted 17 October 2011
Published on the Web: 29 February 2012