

## Genetic variants and risk of prostate cancer using pathway analysis of a genome-wide association study

Y. S. KIM, Y. KIM, J. W. CHOI, H. E. OH, J. H. LEE\*

Department of Pathology, Korea University Ansan Hospital, Ansan, Republic of Korea Department of Pathology, Korea University Ansan Hospital, 123, Jeokgeum-ro, Danwon-Gu, Ansan-Si, Gyeonggi-Do, Republic of Korea

\*Correspondence: [repath@korea.ac.kr](mailto:repath@korea.ac.kr)

Received January 24, 2016 / Accepted April 6, 2016

This study explored candidate causal single nucleotide polymorphisms (SNPs) to clarify the biological mechanism of prostate cancer (PCa). Identify candidate Causal SNPs and Pathways (ICSNPPathway) analysis was applied using a PCa genome-wide association study (GWAS) dataset that included 473,736 SNPs in 1151 cases of PCa and 1156 controls of European ancestry. Five candidate causal SNPs, three candidate causal genes, and two candidate causal pathways were identified using integrating linkage disequilibrium analysis, functional SNP annotation, and pathway-based analysis. The ICSNPPathway analysis suggested three hypothetical mechanisms of PCa. The first was rs13112390, rs13112358, rs2048074 to *nei-like DNA glycosylase 3 (NEIL3)* gene to damaged DNA binding. The second was rs3087386 to *REV1, DNA directed polymerase (REV1)* gene to damaged DNA binding. The third was rs1063134 to *potassium channel, inwardly rectifying subfamily J, member 4 (KCNJ4)* gene to inward rectifier potassium channel activity.

*Key words:* genome-wide association study, pathway-based analysis, prostate cancer

Prostate cancer (PCa) is the most frequently diagnosed malignancy in men. In the United States, an estimated 220,800 new cases and 27,540 PCa-related deaths were projected to occur in 2015 [1]. PCa has a variable clinical course and remarkable genetic heterogeneity [2]. Previous genome-wide association studies (GWASs) of PCa patients identified more than 50 single nucleotide polymorphisms (SNPs) associated with susceptibility for PCa [3].

GWAS has been proven as a valuable tool for identifying the common alleles that confer susceptibility to complex diseases [4, 5]. Increased numbers of GWASs have led to the discovery of novel disease associated genes [4, 5]. One of the advantages of the GWAS approach is that it can reveal causal genes not previously suspected in disease etiology, as well as identify genetic effects of nongenetic DNA region [5]. Improving the results of GWAS concerning the mechanism of disease development and genetic pathways has been a key challenge in interpreting GWAS data [6, 7], because the functional pathway using genomic database has a high statistical power in identifying novel biological pathways that drive disease development [6].

To identify candidate Causal Single Nucleotide polymorphisms and Pathways (ICSNPPathway) analysis was developed

to detect candidate SNPs and their corresponding candidate biological pathways from GWAS data, together with integrating linkage disequilibrium (LD) analysis, functional SNP annotation, and pathway-based analysis (PBA) [8]. This pathway analysis may advance GWAS data interpretation from variants to biological mechanisms, because it distinguishes candidate SNPs and their corresponding candidate pathways [8].

In this study, we performed ICSNPPathway analysis using a PCa GWAS dataset available online, to identify candidate SNPs and promising biological pathways that contribute to PCa susceptibility.

### Materials and methods

**Study population.** The PCa GWAS dataset is publicly available from the NCBI dbGap ([http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000207.v1.p1](http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000207.v1.p1)). This GWAS dataset includes genotypes of 473,736 SNPs obtained from Illumina HumanHap 300 and HumanHap 240 genotyping assays. The study population included 1,151 PCa patients and 1,156 control participants from the Cancer Genetic Markers of Susceptibility (CGEMS) data of prostate

adenocarcinoma [9]. The randomized study population enrollment occurred between 1993 and 2001. The CGEMS cohort consisted of White, non-Hispanic males with no prior history of PCa before randomization [9]. Controls were selected by incidence-density sampling. Control subjects were characterized by the same year of entry into the cohort as the case set, the same 5-year age-at-entry interval (55-59, 60-64, 65-69, 70-74) as the case set, and were observed through the same year of follow-up as the case set with no PCa diagnosis [9]. The dataset was filtered to remove genotyping errors. SNPs were excluded if they displayed significant deviation from Hardy-Weinberg equilibrium in the controls ( $P < 10^{-4}$ ) or had a call rate  $< 90\%$  in cases or controls. Finally, 472,982 SNPs were included for downstream pathway analysis.

**Identification of candidate causal SNPs and pathways.** We conducted ICSNPathway analysis as described previously [10]. Briefly, candidate causal SNPs are pre-selected by LD analysis and the most significant functional SNPs are annotated. This method led to the discovery of biological pathways and mechanisms for the pre-selected candidate causal SNPs using PBA [8]. A full list of PCa GWAS SNP  $P$ -values was entered into the ICSNPathway analysis. ICSNPathway analysis is based on LD analysis and the detection of functional SNPs using improved-gene set enrichment analysis ( $i$ -GSEA). SNPs in LD are identified and the most significant SNPs of a GWAS dataset are identified and analyzed to find more possible candidate causal SNPs based on the extended data set, such as HapMap data. We selected the following optional parameters for LD: CEPH/CEU (Utah residents with ancestry from Northern and Western Europe), a cutoff for LD measurement as  $r^2$  of 0.8, and the maximum distance to search LD neighborhoods as 200kb. The ICSNPathway pre-selects candidate causal SNPs based on functional SNPs (i.e., SNPs that may alter protein, gene expression or the role of protein in the context of the pathway). The functional SNPs include deleterious and non-deleterious non-synonymous SNPs, SNPs leading to a gain or loss of a stop codon, SNPs resulting in a frame shift, SNPs within essential splice sites, and SNPs in regulatory regions. The ICSNPathway server detects pathway-associated traits in the full list of GWAS SNP  $P$ -values using  $i$ -GSEA [8, 10].

“The most significant SNPs” refers to SNPs with a  $P$ -value below a certain threshold. The  $P$ -value threshold to extract the most significant SNPs is specified from the GWAS SNP  $P$ -values. ICSNPathway analysis presents the most significant pathways from the original GWAS when a  $P$ -value threshold  $< 0.01$  is selected. We arranged 2 parameters for the analysis in this study. The first parameter was ‘within gene,’ meaning that only the  $P$ -values of SNPs located within genes were utilized in the PBA algorithm. The second parameter was a false discovery rate (FDR) cutoff of 0.05 for multiple testing corrections. The FDR, which is defined as the expected proportion of false positives among all significant tests, allows us to identify a set of positive candidates. Minimum and maximum cut-off of 5 and 100, respectively, were used to avoid overly narrow or overly broad functional categories [8,10].

We selected 4 pathway databases: the Kyoto Encyclopedia of Genes and Genomes (KEGG) (<http://www.genome.jp/kegg/pathway.html>) [11], the BioCarta (<http://www.biocarta.com/genes/index.asp>), the Gene ontology of biological process (<http://www.geneontology.org>) [12], and the Gene ontology of molecular function (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>) to ensure a comprehensive coverage of pathways, and to obtain high-quality information for well-defined pathways. When a candidate SNP was not present on a particular genotyping array, proxy SNPs in LD with that candidate SNP were identified, based on observed LD patterns in HapMap data (<http://www.broadinstitute.org/mpg/snap>). An SNP annotation and proxy search was carried out to identify and annotate proxy SNPs, using HapMap.

## Results

Utilizing GWAS SNP  $p$ -values as input, ICSNPathway analysis identified five candidate causal SNPs, three candidate causal genes, and two candidate causal pathways ([http://ICSNPathway.psych.ac.cn/getResult.do?tag=A5F58141EFE285F7FD316198EC72E58C\\_1449375126529](http://ICSNPathway.psych.ac.cn/getResult.do?tag=A5F58141EFE285F7FD316198EC72E58C_1449375126529)) (Table 1 and 2; Figure 1). SNP rs13112390, rs13112358, rs2048074, rs3087386, and rs1063134, which were rep-

**Table 1. Candidate causal SNPs**

Candidate causal SNP	Functional class	Gene	Candidate causal pathway*	$-\log_{10}(P)^\dagger$	In LD with	$r^2$	D'	$-\log_{10}(P)^\ddagger$	Chromosome No. Position	RefSNP Alleles
rs13112390	non-synonymous coding	NEIL3	1	2.859	rs13112390	-	-	2.859	4 177353681	A/C
rs13112358	non-synonymous coding	NEIL3	1	2.305	rs13112358	-	-	2.305	4 177353596	C/T
rs2048074	frameshift coding	NEIL3	1	2.244	rs2048074	-	-	2.244	4 177353411	A/G
rs3087386	non-synonymous coding	REV1	1	2.077	rs3087386	-	-	2.077	2 99439044	C/T
rs1063134	regulatory region	KCNJ4	2	2.090	rs1063134	-	-	2.090	22 38426693	C/T

SNP – single nucleotide polymorphism; LD – linkage disequilibrium; \* – numbers indicates the indexes of pathway ranked significance (false discovery rate); No. – number; † –  $-\log_{10}(P)$  values of candidate causal SNPs in the original genome wide association studies (GWAS), -; this SNP is not represented in the original GWAS; ‡ –  $-\log_{10}(P)$  values of SNPs in LD with candidate causal SNPs in the original GWAS

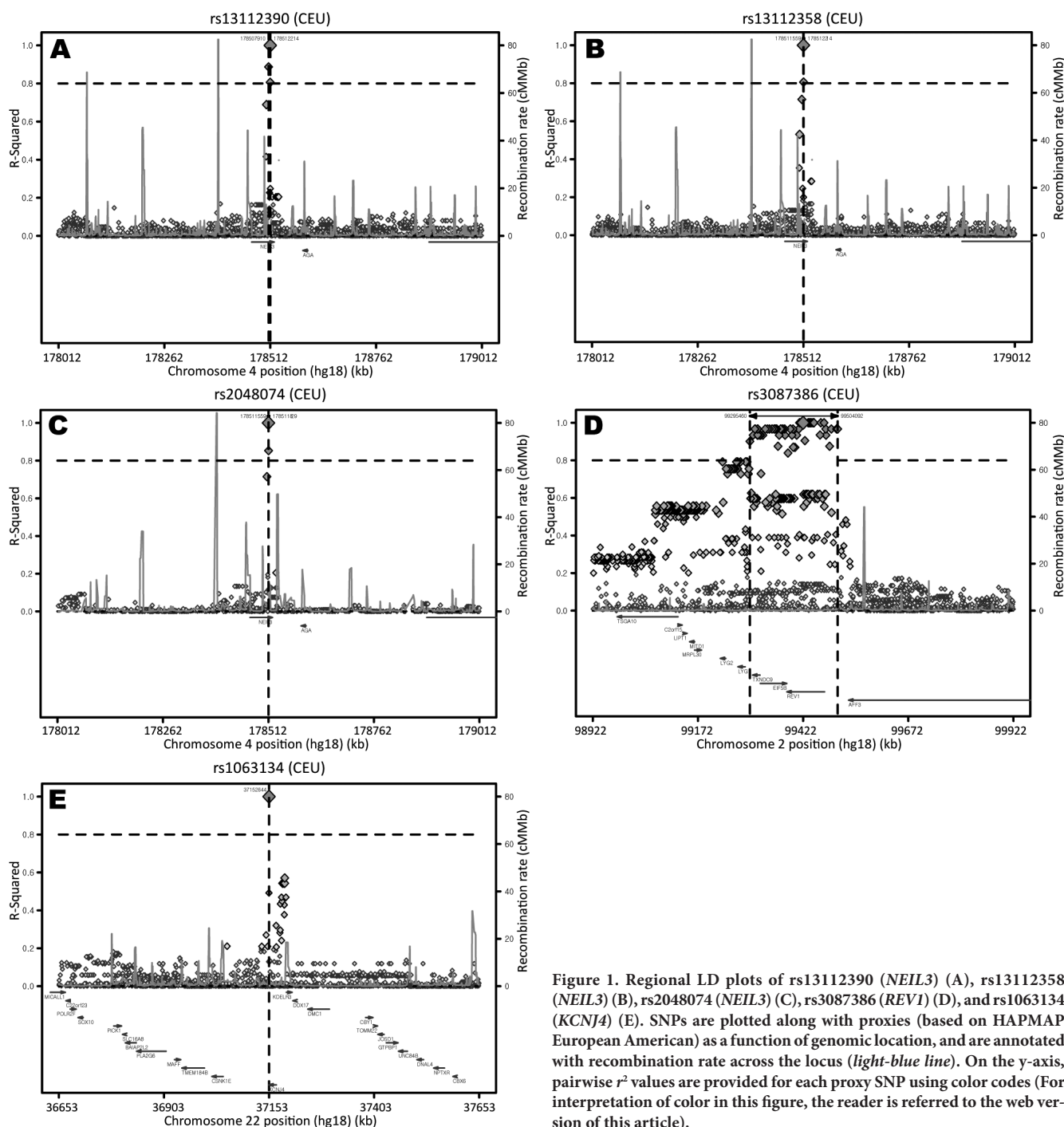


Figure 1. Regional LD plots of rs13112390 (*NEIL3*) (A), rs13112358 (*NEIL3*) (B), rs2048074 (*NEIL3*) (C), rs3087386 (*REV1*) (D), and rs1063134 (*KCNJ4*) (E). SNPs are plotted along with proxies (based on HAPMAP European American) as a function of genomic location, and are annotated with recombination rate across the locus (light-blue line). On the y-axis, pairwise  $r^2$  values are provided for each proxy SNP using color codes (For interpretation of color in this figure, the reader is referred to the web version of this article).

resented in the original GWAS ( $-\log_{10}(P)$ =2.859, 2.305, 2.244, 2.077, and 2.090, respectively), were not in LD with any SNP.

These five candidate causal SNPs, three candidate causal genes, and two causal pathways provided three hypothetical biological mechanisms for PCa. The first was rs13112390, rs13112358 (non-synonymous codings), and rs2048074

(frameshift coding) to *nei-like DNA glycosylase 3 (NEIL3)* gene to damaged DNA binding. The second was rs3087386 (non-synonymous coding) to *REV1, DNA directed polymerase (REV1)* gene to damaged DNA binding. The third was rs1063134 (regulatory region) to *potassium channel, inwardly rectifying subfamily J, member 4 (KCNJ4)* gene to inward rectifier potassium channel activity (Tables 1 and 2).

**Table 2. Candidate causal pathways**

Index	Candidate causal pathway	Description	Nominal <i>P</i>	FDR
1	Damaged DNA binding	Interacting selectively with damaged DNA. Maturity onset diabetes of the young	<0.001	0.031
2	Inward rectifier potassium channel activity	An inwardly rectifying current-voltage relation is one where, at any given driving force, the inward flow of K <sup>+</sup> ions exceeds the outward flow for the opposite driving force. The inward-rectification is due to a voltage-dependent block of the channel pore by a specific ligand or ligands, and, as a result, the macroscopic conductance depends on the difference between the membrane voltage and the K <sup>+</sup> equilibrium potential, rather than on the membrane voltage itself.	<0.001	0.040

FDR; false discovery rate

## Discussion

In this pathway analysis, we identified five candidate causal SNPs, three candidate causal genes, and two candidate causal pathways. These candidate SNPs and pathways provided three hypothetical biological mechanisms. In this genome wide search for pathway associated with PCa, the most strongly associated pathway concerned damaged DNA binding.

The *nei-like DNA glycosylase 3 (NEIL3)* gene belongs to a class of DNA glycosylases that recognizes and removes oxidized DNA bases, and initiates the base excision repair pathway [13]. *NEIL3* is highly expressed in various human cancer cells and tissues, including primary malignant melanomas associated with metastasis [13, 14]. It has been proposed that *NEIL3* may be a potential tumor suppressor gene of hepatocellular carcinoma [15]. In addition, previous studies reported that *NEIL3* SNP rs12645561 was associated with the risk of developing glioma and glioblastoma [16, 17]. These observations may suggest that *NEIL3* is required for the maintenance of cancer cell growth or the progression of malignancy, and that it may play a role in DNA repair as a tumor suppressor gene. In contrast, a conflicting association between *NEIL3* SNP rs1983132 and PCa risk has been reported [18]. Our pathway analysis proposes that the *NEIL3* gene and damaged DNA binding pathway play an important role in PCa susceptibility.

Until now, the association between *NEIL3* gene and PCa patients' survival outcome has not been studied. Hildrestrand *et al.* [19] reported that the cDNA of PCa displayed a higher expression of *NEIL3* than the corresponding normal prostate tissues. Unfortunately, this pathway analysis could not offer a clinical significance of the *NEIL3* gene due to restrictions of the GWAS data. Further studies should clinically investigate the role of the *NEIL3* gene in PCa.

The current pathway analysis identified the candidate SNP rs3087386 within the *REV1*, *DNA directed polymerase (REV1)* gene at chromosome 2q11.1-11.2. *REV1* is an essential regulator of translesion DNA synthesis polymerase, essential for action against damaged bases [20]. It is known that an imbalance of translesion DNA synthesis polymerase leads to genomic instability, which can result in cancer [20]. The SNP rs3087386 has been related to lung cancer risk and

survival in African Americans [21] and Chinese populations [22]. The SNP rs3087386 is a missense mutation (Phe257Ser) that might affect the functional properties of protein, *REV1*, which is associated with risk of lung and cervical cancers [23, 24]. In contrast, the association between *REV1* polymorphisms and breast cancer risk was not found [25]. Until now, *REV1* in relation to PCa has not been studied. Our analysis suggests that candidate *REV1* and the associated pathway may hold a significant position in carcinogenesis of PCa.

The *potassium channel, inwardly rectifying subfamily J, member 4 (KCNJ4)* gene is a member of the inward rectifier potassium channel family. To the best of our knowledge, there have been no studies regarding the *KCNJ4* gene and its associated pathways, and PCa.

Over 20 PCa GWASs have been previously performed and have demonstrated the existence of more than 50 common variants associated with a risk of PCa, including *β-microseminoprotein (MSMB)*, *LOC727677*, and *HNF1B* [3]. Although individual GWASs have been successful in detecting new susceptibility genes for various complex diseases, not all of the GWAS data has been fully analyzed [7]. Pathway analysis of genomic data through functionally derived gene sets can be a powerful analytical tool to reveal previously undiscovered genes or SNPs during gene- or SNP-based analyses [6, 7].

The current ICSNPathway analysis has several limitations. Firstly, the incomplete annotation of the human genome may weaken the pathway-based approach. Secondly, further large scale replication studies are required to establish candidate SNPs, genes, and associated pathways [26]; however, the verification of candidate SNPs in independent datasets was beyond the scope of this pathway analysis. Pathway analysis using GWAS plays a complementary role to the searching of novel genes that confer disease susceptibility [6, 8]. Thirdly, our pathway analysis did not detect the SNPs reported in prior GWASs, such as *MSMB*, *LOC727677*, and *HNF1B*. These discordances may be due to ethnic diversity, differences in sample size or type, and variable GWAS array chips.

ICSNPathway analysis should be performed based on individual GWAS data of only a single ancestry. As such, pathway analysis of Asian or African ancestry will broaden the novel biological pathways of PCa. Interestingly, Ge *et al.* [27] demonstrated the candidate SNPs, genes and pathways

of regulating serum prostate-specific antigen (PSA) levels. At present, pathway analysis does not have the capacity to carry out analysis of combined tasks, such as PSA level regulation and PCa occurrence. Recent rapid developments of bioinformatics tools will allow the analysis of more complex processes in the future.

The choice of an appropriate threshold for significant SNPs can be rather arbitrary [6,28]. The observed SNP level threshold in GWASs has ranged from  $P < 0.05$  to  $P < 5 \times 10^{-8}$  [6]. Different levels of significant cut-off values ( $< 0.01$ ,  $< 0.001$ , or  $< 0.0001$ ) for pathway analysis were assigned, with little difference in the discovery of significant pathways [28]. We adopted the  $P$  value of 0.01, considering the number of significant SNPs that could be entered for the pathway analysis.

In summary, we performed pathway analysis using a PCa GWAS dataset to estimate genetic associations of PCa with the SNPs and pathway levels. We identified 5 candidate causal SNPs, 3 candidate causal genes, 2 candidate causal pathways, and 3 hypothetical biological mechanisms, which may contribute to PCa susceptibility.

**Acknowledgements:** The authors gratefully acknowledge investigators for sharing their valuable GWAS data. This study was supported by grants from Korea University (grant number: K1507981).

## References

- [1] SIEGEL RL, MILLER KD, JEMAL A. Cancer statistics, 2015. *CA Cancer J Clin* 2015; 65: 5–29. <http://dx.doi.org/10.3322/caac.21254>
- [2] BARBIERI CE, TOMLINS SA. Reprint of: The prostate cancer genome: Perspectives and potential. *Urol Oncol* 2015; 33: 95–102. <http://dx.doi.org/10.1016/j.urolonc.2015.01.002>
- [3] CHEN R, REN S, SUN Y. Genome-wide association studies on prostate cancer: the end or the beginning? *Protein Cell* 2013; 4: 677–686. <http://dx.doi.org/10.1007/s13238-013-3055-4>
- [4] HARDY J, SINGLETON A. Genomewide association studies and human disease. *N Engl J Med* 2009; 360: 1759–1768. <http://dx.doi.org/10.1056/NEJMra0808700>
- [5] STRANGER BE, STAHL EA, RAJ T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 2011; 187: 367–383. <http://dx.doi.org/10.1534/genetics.110.120907>
- [6] RAMANAN VK, SHEN L, MOORE JH, SAYKIN AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* 2012; 28: 323–332. <http://dx.doi.org/10.1016/j.tig.2012.03.004>
- [7] ELBERS CC, VAN EIJK KR, FRANKE L, MULDER F, VAN DER SCHOUW YT et al. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol* 2009; 33: 419–431. <http://dx.doi.org/10.1002/gepi.20395>
- [8] ZHANG K, CHANG S, CUI S, GUO L, ZHANG L et al. ICS-Pathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. *Nucleic Acids Res* 2011; 39 (Web Server issue): W437–443.
- [9] YEAGER M, ORR N, HAYES RB, JACOBS KB, KRAFT P et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007; 39: 645–649. <http://dx.doi.org/10.1038/ng2022>
- [10] LEE JH, KIM Y, CHOI JW, KIM YS. Genetic variants and risk of gastric cancer: a pathway analysis of a genome-wide association study. *Springerplus* 2015; 4: 215. <http://dx.doi.org/10.1186/s40064-015-1005-8>
- [11] KANEHISA M, GOTO S, FURUMICHI M, TANABE M, HIRAKAWA M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010; 38 (Database issue): D355–360. <http://dx.doi.org/10.1093/nar/gkp896>
- [12] ASHBURNER M, BALL CA, BLAKE JA, BOTSTEIN D, BUTLER H et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25: 25–29. <http://dx.doi.org/10.1038/75556>
- [13] LIU M, DOUBLIE S, WALLACE SS. Neil3, the final frontier for the DNA glycosylases that recognize oxidative damage. *Mutat Res* 2013; 743–744: 4–11. <http://dx.doi.org/10.1016/j.mrfmmm.2012.12.003>
- [14] KAUFFMANN A, ROSSELLI F, LAZAR V, WINNENPENNINCKX V, MANSUET-LUPO A et al. High expression of DNA repair pathways is associated with metastasis in melanoma patients. *Oncogene* 2008; 27: 565–573. <http://dx.doi.org/10.1038/sj.onc.1210700>
- [15] ZHANG H, MA H, WANG Q, CHEN M, WENG D et al. Analysis of loss of heterozygosity on chromosome 4q in hepatocellular carcinoma using high-throughput SNP array. *Oncol Rep* 2010; 23: 445–455.
- [16] BETHKE L, WEBB E, MURRAY A, SCHOEMAKER M, JOHANSEN C et al. Comprehensive analysis of the role of DNA repair gene polymorphisms on risk of glioma. *Hum Mol Genet* 2008; 17: 800–805. <http://dx.doi.org/10.1093/hmg/ddm351>
- [17] JIN TB, LI XL, YANG H, JIRI M, SHI XG et al. Association of polymorphisms in FLT3, EGFR, ALOX5, and NEIL3 with glioblastoma in the Han Chinese population. *Med Oncol* 2013; 30: 718. <http://dx.doi.org/10.1007/s12032-013-0718-1>
- [18] BARRY KH, KOUTROS S, BERNDT SI, ANDREOTTI G, HOPPIN JA et al. Genetic variation in base excision repair pathway genes, pesticide exposure, and prostate cancer risk. *Environ Health Perspect* 2011; 119: 1726–1732. <http://dx.doi.org/10.1289/ehp.1103454>
- [19] HILDRESTRAND GA, NEURAUER CG, DIEP DB, CASSELLANOS CG, KRAUSS S et al. Expression patterns of Neil3 during embryonic brain development and neoplasia. *BMC Neurosci* 2009; 10: 45. <http://dx.doi.org/10.1186/1471-2202-10-45>
- [20] SHARMA S, HELCHOWSKI CM, CANMAN CE. The roles of DNA polymerase zeta and the Y family DNA polymerases in promoting or preventing genome instability. *Mutat Res* 2013; 743–744: 97–110. <http://dx.doi.org/10.1016/j.mrfmmm.2012.11.002>
- [21] SPITZ MR, AMOS CI, LAND S, WU X, DONG Q et al. Role of selected genetic variants in lung cancer risk in African Americans. *J Thorac Oncol* 2013; 8: 391–397. <http://dx.doi.org/10.1097/JTO.0b013e318283da29>

- [22] XU HL, GAO XR, ZHANG W, CHENG JR, TAN YT et al. Effects of polymorphisms in translesion DNA synthesis genes on lung cancer risk and prognosis in Chinese men. *Cancer Epidemiol* 2013; 37: 917–922. <http://dx.doi.org/10.1016/j.canep.2013.08.003>
- [23] SAKIYAMA T, KOHNO T, MIMAKI S, OHTA T, YAN-AGITANI N et al. Association of amino acid substitution polymorphisms in DNA repair genes TP53, POLI, REV1 and LIG4 with lung cancer risk. *Int J Cancer* 2005; 114: 730–737. <http://dx.doi.org/10.1002/ijc.20790>
- [24] HEX X, YE F, ZHANG J, CHENG Q, SHEN J et al. REV1 genetic variants associated with the risk of cervical carcinoma. *Eur J Epidemiol* 2008; 23: 403–409. <http://dx.doi.org/10.1007/s10654-008-9251-5>
- [25] VARADI V, BEVIER M, GRZYBOWSKA E, JOHANSSON R, ENQUIST K et al. Genetic variation in genes encoding for polymerase zeta subunits associates with breast cancer risk, tumour characteristics and survival. *Breast Cancer Res Treat* 2011; 129: 235–245. <http://dx.doi.org/10.1007/s10549-011-1460-z>
- [26] JIA P, WANG L, MELTZER HY, ZHAO Z. Pathway-based analysis of GWAS datasets: effective but caution required. *Int J Neuropsychopharmacol* 2011; 14: 567–572. <http://dx.doi.org/10.1017/S1461145710001446>
- [27] GE YZ, XU Z, XU LW, YU P, ZHAO Y et al. Pathway analysis of genome-wide association study on serum prostate-specific antigen levels. *Gene* 2014; 551: 86–91. <http://dx.doi.org/10.1016/j.gene.2014.08.044>
- [28] LAMBERT JC, GRENIER-BOLEY B, CHOURAKI V, HEATH S, ZELENKA D et al. Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J Alzheimers Dis* 2010; 20: 1107–1118.