nature genetics

A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci

Jin-Xin Bei[1,2,9], Yi Li[3,9], Wei-Hua Jia[1,2], Bing-Jian Feng[1,2,4], Gangqiao Zhou[5], Li-Zhen Chen[1,2], Qi-Sheng Feng[1,2], Hui-Qi Low[3], Hongxing Zhang[5], Fuchu He[5], E Shyong Tai[6,7], Tiebang Kang[1,2], Edison T Liu[8], Jianjun Liu[1,3,10] & Yi-Xin Zeng[1,2,10]

To identify genetic susceptibility loci for nasopharyngeal carcinoma (NPC), a genome-wide association study was performed using 464,328 autosomal SNPs in 1,583 NPC affected individuals (cases) and 1,894 controls of southern Chinese descent. The top 49 SNPs from the genome-wide association study were genotyped in 3,507 cases and 3,063 controls of southern Chinese descent from Guangdong and Guangxi. The seven supportive SNPs were further confirmed by transmission disequilibrium test analysis in 279 trios from Guangdong. We identified three new susceptibility loci, *TNFRSF19* on 13q12 (rs9510787, $P_{combined} = 1.53 \times 10^{-9}$, odds ratio (OR) = 1.20), *MDS1-EVI1* on 3q26 (rs6774494, $P_{combined} = 1.34 \times 10^{-8}$, OR = 0.84) and the *CDKN2A-CDKN2B* gene cluster on 9p21 (rs1412829, $P_{combined} = 4.84 \times 10^{-7}$, OR = 0.78). Furthermore, we confirmed the role of *HLA* by revealing independent associations at rs2860580 ($P_{combined} = 4.88 \times 10^{-67}$, OR = 0.58), rs2894207 ($P_{combined} = 3.42 \times 10^{-33}$, OR = 0.61) and rs28421666 ($P_{combined} = 2.49 \times 10^{-18}$, OR = 0.67). Our findings provide new insights into the pathogenesis of NPC by highlighting the involvement of pathways related to *TNFRSF19* and *MDS1-EVI1* in addition to *HLA* molecules.

NPC is a squamous-cell carcinoma that arises in the epithelial lining of the nasopharynx[1]. This neoplasm has remarkable ethnic and geographic distribution, with a high prevalence in southern China, southeast Asia, northern Africa and Alaska[2]. The annual incidence rate reaches 25 cases per 100,000 people in the endemic regions, which is about 25-fold higher than that in the rest of the world[2,3]. Familial clustering of NPC has been observed in diverse populations[4]. Elevated levels of circulating free Epstein-Barr virus (EBV) DNA and the EBV-related antibodies in sera, as well as clonal EBV DNA in tumor cells, were consistently detected in individuals with NPC[1,5]. Furthermore, consumption of salted and pickled foods and smoking have also been demonstrated to increase the risk of NPC[2]. These studies revealed that NPC is a multifactorial malignancy associated with EBV infection and is influenced by both genetic and environmental factors[1,6].

The association of *HLA* (encoding human leukocyte antigen) subtypes with NPC susceptibility has been extensively studied (**Supplementary Table 1**), and *HLA-A* has been consistently shown to be associated with NPC in both association studies (**Supplementary Table 1**) and linkage studies[7,8]. We and others have also reported additional susceptibility loci on 4p15.1–q12 (ref. 9), 3p21 (ref. 10) and 5p13 (ref. 11) through linkage studies of NPC families from southern China. In addition, many candidate genes have also been implicated in NPC susceptibility (**Supplementary Table 2**). Recently, a genome-wide association study (GWAS) of NPC in Taiwanese individuals reported the association of *HLA-A* and an independent association located between *GABBR1* and *HLA-F*[12]. Another GWAS of NPC was performed in Malaysian Chinese subjects and reported an association at *ITGA9* on 3p21 (ref. 13).

Here, we conducted a large GWAS of NPC in southern Chinese individuals by genotyping 620,901 SNPs in 1,615 cases and 1,025 controls of Han Chinese descent from Guangdong and an additional 1,008 Singapore Chinese controls, who share the same ancestral origin with Han Chinese individuals in southern China[14]. After stringent quality-control filtering (see Online Methods for details), 464,328 autosomal SNPs in 1,583 cases and 1,894 controls (972 Guangdong subjects and 922 Singapore subjects) were retained for statistical analysis. Principal component analysis (PCA) showed that the 1,583 cases and 972 Guangdong controls were genetically well matched, but the inclusion of the 922 Singapore controls caused moderate population stratification in the GWAS sample (**Supplementary Fig. 1**). We therefore performed genotype-phenotype association analysis using the Cochran-Armitage trend test with PCA-based correction for population stratification and with adjustment for age and gender. A quantile-quantile plot revealed a good match between the distributions of the observed *P* values and those expected by chance, except for the presence of a strong deviation within the upper tail of the distribution

[1]State Key Laboratory of Oncology in Southern China, Guangzhou, China. [2]Department of Experimental Research, Sun Yat-sen University Cancer Center, Guangzhou, China. [3]Human Genetics, Genome Institute of Singapore, A*STAR, Singapore. [4]Department of Dermatology, University of Utah School of Medicine, Salt Lake City, Utah, USA. [5]The State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing, China. [6]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore. [7]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore. [8]Cancer Biology, Genome Institute of Singapore, A*STAR, Singapore. [9]These authors contributed equally to the work. [10]These authors jointly directed this work. Correspondence should be addressed to Y.-X.Z. (zengyx@sysucc.org.cn) or J.L. (liuj3@gis.a-star.edu.sg).

Received 6 January; accepted 6 May; published online 30 May 2010; doi:10.1038/ng.601

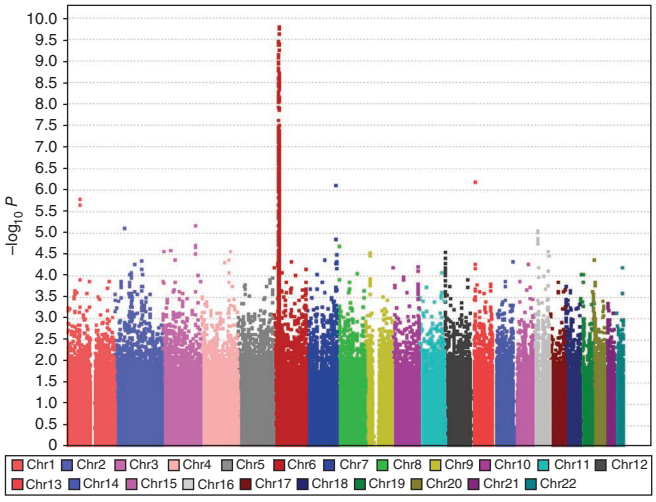

Figure 1 Manhattan plot of the genome-wide *P* values of association. Association was assessed using Cochran-Armitage trend test in logistic regression analysis with adjustment for age, gender and the top ten principal components of population stratification. The $-\log_{10} P$ values (*y* axis) of 464,328 SNPs in 1,583 NPC cases and 1,894 controls are presented based on their chromosomal positions (*x* axis). The points with $P < 10^{-10}$ were truncated, and the smallest *P* value is 1.34×10^{-28}.

(**Supplementary Fig. 2**). A small genomic-control inflation factor (λ_{gc}) of 1.03 (calculated excluding the *MHC* SNPs) indicates a minimal inflation of the observed genome-wide association significance due to population stratification.

The GWAS revealed multiple associations within the *MHC* region and suggestive associations on chromosomes 1, 3, 7 and 13 ($P < 10^{-5}$) (**Fig. 1**). Forward conditional logistic regression analysis was performed on 3,943 *MHC* SNPs, which revealed associations at rs2860580, rs2894207 and rs28421666 (**Supplementary Fig. 3**). However, we did not observe any association within the *ITGA9* locus[13], the linkage loci on 4p15.1–q12, 3p21 and 5p13 (refs. 9–11), or other previously reported candidate genes and regions (**Supplementary Fig. 4** and **Supplementary Table 3**).

To validate these findings, we genotyped the top 49 SNPs from the GWAS in two independent Chinese samples from southern China, including 2,737 cases and 2,050 controls from Guangdong and 770 cases and 1,013 controls from Guangxi. Six SNPs showed significant association in the combined validation sample, including three *HLA* SNPs (rs2860580, rs2894207 and rs28421666;$P < 10^{-10}$) and three non-*HLA* SNPs (rs6774494, rs1412829 and rs9510787; $P < 10^{-3}$) (**Table 1**). rs1572072 also showed consistent association (OR = 0.88, $P = 1.88 \times 10^{-3}$), but the association did not achieve statistical significance after correction for testing 49 SNPs. All seven SNPs showed consistent association between the two independent validation samples, although the associations at rs28421666, rs9510787 and rs1572072 did not reach statistical significance in the Guangxi sample, likely due to the smaller sample size of that cohort (**Supplementary Table 4**). In the combined GWAS and validation case-control samples, the associations at rs2860580, rs2894207, rs28421666, rs6774494 and rs9510787 surpassed genome-wide significance ($P < 5 \times 10^{-8}$) (**Table 1**). The remaining 42 SNPs were not validated (**Supplementary Table 5**).

As further confirmation, the seven associated SNPs were subjected to transmission disequilibrium test (TDT) analysis in a third independent sample of 279 family trios from Guangdong. The TDT analysis revealed associations at rs2860580 ($P = 3.37 \times 10^{-4}$), rs2894207 ($P = 9.32 \times 10^{-4}$), rs6774494 ($P = 4.8 \times 10^{-2}$), rs9510787 ($P = 2.6 \times 10^{-2}$) and rs1572072 ($P = 6.72 \times 10^{-4}$) (**Table 1**). It also revealed consistent associations at rs28421666 (OR = 0.78) and rs1412829 (OR = 0.79), although the evidence was not statistically significant, likely due to the low allele frequencies of these two SNPs (**Table 1**). The combined case-control and trio samples revealed genome-wide significant associations at rs2860580 ($P_{combined} = 4.88 \times 10^{-67}$, OR = 0.58), rs2894207 ($P_{combined} = 3.42 \times 10^{-33}$, OR = 0.61) and rs28421666 ($P_{combined} = 2.49 \times 10^{-18}$, OR = 0.67) within the *HLA* region, rs9510787 ($P_{combined} = 1.53 \times 10^{-9}$, OR = 1.20) and rs1572072 ($P_{combined} = 1.30 \times 10^{-8}$, OR = 0.84) on 13q12, and rs6774494 ($P_{combined} = 1.34 \times 10^{-8}$, OR = 0.84) on 3q26 (**Table 1**). Our study also revealed an association at rs1412829 on 9p21 ($P_{combined} = 4.84 \times 10^{-7}$, OR = 0.78) which was just below the threshold for genome-wide significance (**Table 1**). Furthermore, in the combined case-control sample, we did not detect any interaction between the three *HLA* SNPs and four non-*MHC* SNPs, or among the four non-*MHC* SNPs (**Supplementary Tables 6** and **7**).

We further investigated the impact of the Singapore Chinese controls on our association findings. First, the seven associated SNPs have similar allele frequencies in the 922 Singapore Chinese controls and 972 Guangdong Chinese controls, except in the case of rs2894207 within *HLA*, where a significant difference in allele frequencies was observed between the two samples (18% compared to 14%, $P = 0.006$; **Supplementary Table 8**). The difference, however, had a minor impact on the association at rs2894207 as shown by the fact that the strength of the association remained the same (OR = 0.61, $P = 7.12 \times 10^{-32}$) after removing the Singapore Chinese controls from the analysis. Similarly, the removal of the Singapore Chinese controls from the analysis also had a minor impact on the associations at the other six SNPs (**Supplementary Table 9**).

The associations at rs1572072 and rs9510787 on 13q12 were within a 200-kb region of high linkage disequilibrium (LD) containing only *TNFRSF19* (encoding the tumor necrosis factor (TNF) receptor superfamily, member 19) (**Fig. 2a**). Imputation analysis of the area surrounding *TNFRSF19* in the GWAS sample revealed additional associations within the same high-LD region (**Fig. 2a**). rs1572072 and rs9510787 are 17 kb upstream and are within the fifth intron of *TNFRSF19*, respectively. These two SNPs are separated by a recombination hotspot (**Fig. 2a**) and are not correlated ($r^2 = 0.07$ and $D' = 0.60$; **Supplementary Table 10**). The conditional association analysis in the combined case-control sample showed that these two SNPs were not significant at the genome-wide level after adjustment for each other's effect and that the strengths of the associations (that is, the ORs) were only slightly changed after this adjustment (**Supplementary Table 11**). The haplotype analysis of the two SNPs, however, revealed a single associated haplotype (**Supplementary Table 12**). Our results suggest that the two SNPs might tag the same causal variant but that analyzing both SNPs, rather than only the top SNP rs9510787, will better assess the risk associated with this locus.

The association at rs6774494 on 3q26 was within a high-LD region of about 50 kb covering *MDS1-EVI1* (encoding myelodysplasia 1 and ecotropic viral insertion site 1 fusion proteins) (**Fig. 2b**). The imputation analysis revealed additional associations, but all were within the same high-LD region. The suggestive association at rs1412829 on 9p21 was within a high-LD region of ~187 kb where *CDKN2A*, *CDKN2B* and *CDKN2BAS* loci are located (**Fig. 2c**). Therefore, our study identified two new susceptibility loci, at *TNFRSF19* and *MDS1-EVI1*, with genome-wide significant association and one, at *CDKN2A-CDKN2B*, with suggestive association.

Our study revealed associations within the *HLA* region at rs2860580, located 4 kb upstream of *HLA-A*, rs2894207, located

Table 1 Association results of seven SNPs in the GWAS, case-control and trio validation samples and in the combined samples

				Cases and controls								Families			
				GWAS[a]		Validation 1[b]		Combined GWAS and validation 1[c]		Validation 2[d]		All of the samples[e]		Combined results	
SNP	Chr.	Locus	MA	MAF[f]	*P*	MAF[g]	*P*	OR (95% CI)	*P*	OR (95% CI)	TR/UTR	TDT *P*	OR (95% CI)	*P*	OR (95% CI)
rs2860580	6	*HLA-A*	A	0.26/0.39/0.37	1.34×10^{-28}	0.27/0.38	1.83×10^{-38}	0.60 (0.56–0.65)	3.65×10^{-65}	0.58 (0.54–0.62)	94/150	3.37×10^{-4}	0.63 (0.48–0.81)	4.88×10^{-67}	0.58 (0.55–0.62)
rs2894207	6	*HLA-B/C*	G	0.10/0.18/0.14	1.22×10^{-16}	0.12/0.18	6.89×10^{-17}	0.66 (0.60–0.73)	1.83×10^{-31}	0.61 (0.57–0.67)	53/93	9.32×10^{-4}	0.57 (0.41–0.80)	3.42×10^{-33}	0.61 (0.57–0.66)
rs28421666	6	*HLA-DQ/DR*	G	0.10/0.15/0.11	3.54×10^{-9}	0.09/0.12	9.48×10^{-11}	0.68 (0.61–0.77)	1.40×10^{-18}	0.66 (0.60–0.72)	56/72	0.16	0.78 (0.55–1.10)	2.49×10^{-18}	0.67 (0.61–0.73)
rs9510787	13	*TNFRSF19*	G	0.40/0.35/0.30	6.32×10^{-7}	0.39/0.37	4.29×10^{-4}	1.14 (1.06–1.22)	9.57×10^{-9}	1.20 (1.10–1.30)	149/113	2.60×10^{-2}	1.32 (1.03–1.68)	1.53×10^{-9}	1.20 (1.10–1.30)
rs1572072	13	*TNFRSF19*	A	0.24/0.27/0.29	3.52×10^{-4}	0.24/0.26	1.88×10^{-3}	0.88 (0.81–0.95)	3.47×10^{-6}	0.86 (0.80–0.91)	72/119	6.72×10^{-4}	0.61 (0.45–0.81)	1.30×10^{-8}	0.84 (0.79–0.90)
rs6774494	3	*MDS1-EVI1*	G	0.32/0.37/0.39	6.53×10^{-6}	0.32/0.35	5.90×10^{-4}	0.88 (0.82–0.95)	5.05×10^{-8}	0.85 (0.80–0.90)	100/130	4.80×10^{-2}	0.77 (0.59–0.99)	1.34×10^{-8}	0.84 (0.79–0.89)
rs1412829	9	*CDKN2A/2B*	G	0.08/0.12/0.10	2.78×10^{-5}	0.09/0.11	8.16×10^{-4}	0.82 (0.73–0.92)	3.51×10^{-7}	0.78 (0.71–0.86)	33/42	0.3	0.79 (0.5–1.24)	4.84×10^{-7}	0.78 (0.71–0.85)

Chr., chromosome; MA, minor allele; MAF, minor allele frequency; OR, odds ratio for minor allele; and TR/UTR, transmitted/untransmitted counts.
[a]GWAS, *n* = 1,583 cases and 1,894 controls. [b]Validation 1, *n* = 3,507 cases and 3,063 controls. [c]Combined GWAS and validation 1, *n* = 5,090 cases and 4,957 controls. [d]Validation 2, *n* = 279 trios. [e]All of the samples, *n* = 5,090 cases, 4,957 controls and 279 trios.
[f]MAFs in the cases/Guangdong controls/Singapore Chinese controls. [g]MAFs in the cases/controls.

within *HLA-B–HLA-C*, and rs28421666, located between the *HLA-DR–DQ* loci (**Table 1** and **Supplementary Fig. 3**). The imputation of *HLA-A* alleles in our GWAS sample revealed the strong association of *HLA-A*1101* with a reduced risk for NPC (OR = 0.56, $P = 3 \times 10^{-18}$) as previously reported[12,15]. *HLA-A*1101* and rs2860580 are in strong LD ($r^2 = 0.85$, $D' = 1.0$), and their strengths of association are comparable (see **Supplementary Note**). Further study will be needed to confirm whether *HLA-A*1101* is the causal allele of the association observed at rs2860580. The conditional association analysis in the combined case-control sample revealed that each of the three SNPs remained genome-wide significant after adjusting for the effects of the other two SNPs (**Supplementary Table 13**), which is consistent with the presence of low LD among the three SNPs (**Supplementary Table 10**). Furthermore, all the common haplotypes of the three SNPs(frequency > 5%) showed significant frequency differences between the case and control groups (**Supplementary Table 12**). Taken together, our results suggest multiple independent associations within *HLA*, consistent with previous reports of the associations of certain *HLA-A*, *HLA-B* and *HLA-C* alleles across diverse populations as well as the associations of *HLA-DQB1* and *HLA-DRB1* alleles in Tunisian and Taiwan Chinese populations (see **Supplementary Table 1**). However, further studies are needed to reveal these independent causative variants within *HLA*.

The recent GWAS in Taiwanese individuals reported a strong association at rs2517713 within *HLA-A*[12]. In our study, rs2517713 was not genotyped, but it is located only 11 kb away from rs2860580, our most significant SNP within *HLA*. The two SNPs are in complete LD ($r^2 = 0.99$ and $D' = 1$, based on Han Chinese (CHB) and Japanese (JPT) ancestries from HapMap data). Therefore, the two studies revealed the consistent association within *HLA-A*. The previous study also reported another independent association at rs29232 (OR = 0.60, $P = 8.97 \times 10^{-17}$), located between *GABBR1* and *HLA-F*[12]. rs29232 also showed a similar association in our GWAS sample (OR = 0.64, $P = 3.90 \times 10^{-18}$), but the strength of the association was greatly reduced ($OR_{adjusted} = 0.82$, $P_{adjusted} = 1.39 \times 10^{-3}$) after controlling for the effect of rs2860580. The strength of the association at rs2860580 was also reduced ($OR_{adjusted} = 0.59$, $P_{adjusted} = 1.27 \times 10^{-14}$) after adjusting for rs29232 (**Supplementary Table 14**). Furthermore, rs2860580 and rs29232 are in considerable LD ($r^2 = 0.29$ and $D' = 0.80$, according to HapMap CHB+JPT data). Together, our results suggest that the associations at rs29232 and rs2860580 are correlated, rather than being independent. Nevertheless, both the studies revealed the stronger genetic effect of the *HLA* loci than the non-*HLA* loci, which may partially explain why only the association of *HLA* was consistently discovered by various studies, even in ones using small sample sizes (see **Supplementary Table 1**). Notably, the strong association of *HLA* has also been demonstrated in non-Hodgkin lymphoma[16], another malignancy related to EBV infection, but has not yet been shown to be associated with other cancers (see **Supplementary Table 15**). This may suggest that HLA-mediated immune responses play an important role in virus infection–related cancers. HLA-A11 was shown to present immunodominant EBV epitopes and induce cytotoxic T-lymphocyte responses against EBV-infected cells, which might explain the decreased risk of NPC among individuals with *HLA-A11* (ref. 17). The carriers of other *HLA* risk-associated alleles may also have differential viral antigen–presenting capacity and thus the ability to activate diverse antiviral immune responses, affecting the immune surveillance of viral infection and subsequently NPC susceptibility.

TNFRSF19 encodes a member of the TNF receptor superfamily[18]. When overexpressed, TNFRSF19 activates the c-Jun N-terminal kinase (JNK) pathway and induces caspase-independent cell death[19]. Given the epithelial expression of TNFRSF19 in many embryonic

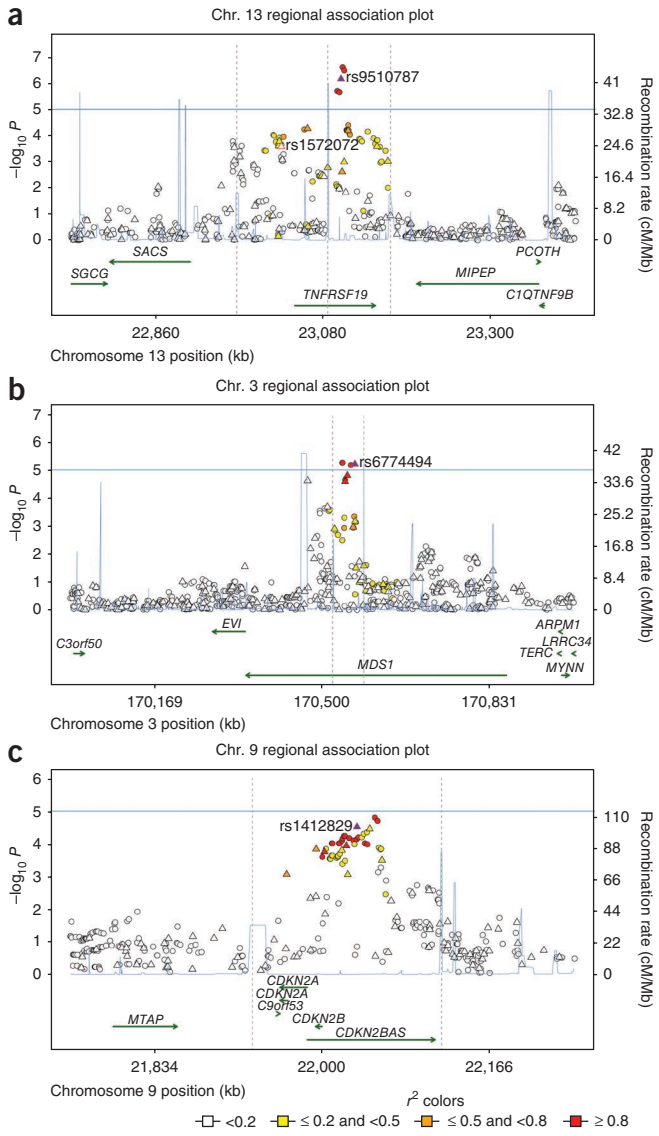

Figure 2 Regional plots of association results and recombination rates within three non-*HLA* susceptibility loci. (**a**–**c**) Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates within the three non-*HLA* susceptibility loci: *TNFRSF19* (**a**), *MDS1-EVI1* (**b**) and *CDKN2A-CDKN2B* (**c**). For each plot, the $-\log_{10}$ *P* values (*y* axis) of the SNPs are presented according to their chromosomal positions (*x* axis), with a blue horizontal line included to indicate suggestive genome-wide significance (10^{-5}). The genetic recombination rates (cM/Mb) (estimated by using the HapMap CHB+JPT samples) are shown with a light blue line, and the genes within the interested region were annotated and shown as green arrows. The top genotyped SNP (labeled by rs ID) is indicated by a red-edged triangle, and the r^2 values of the rest of the SNPs (black-edged triangles) with the top genotyped SNP are indicated by different colors.

tissues[20] and the epithelial origin of NPC, the dysregulation of TNFRSF19 in the epithelium of the nasopharynx may be involved in NPC. Moreover, EBV-encoded latent membrane protein 1 is oncogenic and drives cell transformation through a mechanism similar to the TNF receptor family members[21], lending further biological plausibility to the involvement of *TNFRSF19* in NPC.

MDS1-EVI1 encodes three proteins, EVI1, MDS1 and the fusion protein MDS1-EVI1 (ref. 22). EVI1 is a transcription factor involved in leukemic transformation of hematopoietic cells[22]. EVI1 can suppress the effect of transforming growth factor (TGF)-β on growth inhibition, which in turn promotes tumor growth; EVI1 can also protect cells from stress-induced cell death by inhibiting c-JNK[22,23]. In contrast, when EVI1 was fusioned with MDS1, its capacity to repress TGF-β signaling was significantly impaired[24]. Given that TGF-β and JNK signaling pathways are known to be involved in EBV-related tumorigenesis of NPC[25,26], the interruption of the balance between EVI1 and MDS1-EVI1 proteins may be important for the pathogenesis of NPC.

CDKN2A and *CDKN2B* are known tumor suppressor genes whose double deletion predisposes mice to spontaneous and carcinogen-induced cancers, including leukemia[27,28]. In NPC, the homozygous deletion of both genes was detected in about 40% of primary tumors; moreover, tumor suppressor function and promoterhypermethylation of *CDKN2A* were consistently demonstrated in NPC cell lines[29]. Notably, rs1412829, our top SNP within the *CDKN2A-CDKN2B* locus, was also found to be strongly associated with high-grade glioma in European population[30].

Taken together, our identification of *TNFRSF19, MDS1-EVI1* and *CDKN2A-CDKN2B* as susceptibility loci suggest an important role for the TGF-β and JNK signaling pathways in the pathogenesis of NPC. It is also noteworthy that *TNFRSF19*, *MDS1-EVI1* and *CDKN2A-CDKN2B* are all involved in leukemogenesis, suggesting that there could be a partially shared pathogenic mechanism between hematological malignancy and NPC. This possibility is further supported by the elevated incidence rate of leukemia and other hematological malignancies in individuals with NPC[31,32]. Further studies are needed to investigate the interactions between these genetic susceptibility loci and well-established non-genetic risk factors, particularly EBV infection, and their contributions to the endemics of NPC in southern China.

URL. R project, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on theNature Genetics website.

ACKNOWLEDGMENTS

We would like to thank all the subjects and healthy volunteers who participated in this work. We also want to thank the staffs of Bank of Tumor Resource at Sun Yat-sen University Cancer Center (SYSUCC) for help in sample storage. We also want to thank W.-Y. Meah, X. Chen, H.-B. Toh, K.-K. Heng, C.-H. Wong and D.E.K. Tan, who performed the genotyping, and R.T.H. Ong, J. Chen, K.-S. Sim and E. Tantoso for their assistance in data analysis. This work was funded by the High-Tech Research and Development Program of China (863 Plan, 2006AA02A404), the Natural Science Foundation of China (30621063, 30872929 and u0732005), the National Basic Research Program of China (973 Plan; 2004CB518604 and 2006CB910104), Beijing Science & Technology nova program (2006A54), the National Science and Technology Support Program of China (2006BAI02A11) and the Agency for Science, Technology and Research of Singapore.

AUTHOR CONTRIBUTIONS

Y.-X.Z. was the overall study principal investigator who conceived the study and obtained financial support. Y.-X.Z., J.-X.B. and J.L. designed and oversaw the study. W.-H.J., B.-J.F., Q.-S.F. and L.-Z.C. initiated or participated in the recruitment of Guangdong Chinese samples and preparation of biological samples. H.Z., G.Z. and F.H. were responsible for the recruitment and sample preparation of Guangxi Chinese samples. J.-X.B. conducted sample inclusion and data management. B.-J.F. helped to impute the *HLA* alleles. J.-X.B. and Y.L. undertook the statistical analyses under guidance from J.L. and with help from H.-Q.L. J.-X.B. interpreted the results, drafted and synthesized the manuscript. E.T.L., T.K. and E.S.T. reviewed the manuscript and J.L. helped to revise it.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Wei, W.I. & Sham, J.S. Nasopharyngeal carcinoma. *Lancet* **365**, 2041–2054 (2005).
2. McDermott, A.L., Dutt, S.N. & Watkinson, J.C. The aetiology of nasopharyngeal carcinoma. *Clin. Otolaryngol. Allied Sci.* **26**, 82–92 (2001).
3. Yu, M.C. & Yuan, J.M. Epidemiology of nasopharyngeal carcinoma. *Semin. Cancer Biol.* **12**, 421–429 (2002).
4. Zeng, Y.X. & Jia, W.H. Familial nasopharyngeal carcinoma. *Semin. Cancer Biol.* **12**, 443–450 (2002).
5. Li, D.J. *et al.* The dominance of China 1 in the spectrum of Epstein-Barr virus strains from Cantonese patients with nasopharyngeal carcinoma. *J. Med. Virol.* **81**, 1253–1260 (2009).
6. Young, L.S. & Rickinson, A.B. Epstein-Barr virus: 40 years on. *Nat. Rev. Cancer* **4**, 757–768 (2004).
7. Lu, S.J. *et al.* Linkage of a nasopharyngeal carcinoma susceptibility locus to the HLA region. *Nature* **346**, 470–471 (1990).
8. Lu, C.C. *et al.* Nasopharyngeal carcinoma-susceptibility locus is localized to a 132 kb segment containing HLA-A using high-resolution microsatellite mapping. *Int. J. Cancer* **115**, 742–746 (2005).
9. Feng, B.J. *et al.* Genome-wide scan for familial nasopharyngeal carcinoma reveals evidence of linkage to chromosome 4. *Nat. Genet.* **31**, 395–399 (2002).
10. Xiong, W. *et al.* A susceptibility locus at chromosome 3p21 linked to familial nasopharyngeal carcinoma. *Cancer Res.* **64**, 1972–1974 (2004).
11. Hu, L.F. *et al.* A genome-wide scan suggests a susceptibility locus on 5p 13 for nasopharyngeal carcinoma. *Eur. J. Hum. Genet.* **16**, 343–349 (2008).
12. Tse, K.P. *et al.* Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the *HLA* region at chromosome 6p21.3. *Am. J. Hum. Genet.* **85**, 194–203 (2009).
13. Ng, C.C. *et al.* A genome-wide association study identifies *ITGA9* conferring risk of nasopharyngeal carcinoma. *J. Hum. Genet.* **54**, 392–397 (2009).
14. Chen, J. *et al.* Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am. J. Hum. Genet.* **85**, 775–785 (2009).
15. Hildesheim, A. *et al.* Association of *HLA* class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan. *J. Natl. Cancer Inst.* **94**, 1780–1789 (2002).
16. Skibola, C.F. *et al.* Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.* **41**, 873–875 (2009).
17. Rickinson, A.B. & Moss, D.J. Human cytotoxic T lymphocyte responses to Epstein-Barr virus infection. *Annu. Rev. Immunol.* **15**, 405–431 (1997).
18. Hu, S., Tamada, K., Ni, J., Vincenz, C. & Chen, L. Characterization of TNFRSF19, a novel member of the tumor necrosis factor receptor superfamily. *Genomics* **62**, 103–107 (1999).
19. Eby, M.T., Jasmin, A., Kumar, A., Sharma, K. & Chaudhary, P.M. TAJ, a novel member of the tumor necrosis factor receptor family, activates the c-Jun N-terminal kinase pathway and mediates caspase-independent cell death. *J. Biol. Chem.* **275**, 15336–15342 (2000).
20. Morikawa, Y., Hisaoka, T., Kitamura, T. & Senba, E. TROY, a novel member of the tumor necrosis factor receptor superfamily in the central nervous system. *Ann. NY Acad. Sci.* **1126**, A1–A10 (2008).
21. Eliopoulos, A.G. & Young, L.S. LMP1 structure and signal transduction. *Semin. Cancer Biol.* **11**, 435–444 (2001).
22. Métais, J.Y. & Dunbar, C.E. The MDS1–EVI1 gene complex as a retrovirus integration site: impact on behavior of hematopoietic cells and implications for gene therapy. *Mol. Ther.* **16**, 439–449 (2008).
23. Kurokawa, M. *et al.* The evi-1 oncoprotein inhibits c-Jun N-terminal kinase and prevents stress-induced cell death. *EMBO J.* **19**, 2958–2968 (2000).
24. Nitta, E. *et al.* Oligomerization of Evi-1 regulated by the PR domain contributes to recruitment of corepressor CtBP. *Oncogene* **24**, 6165–6173 (2005).
25. Xu, J., Menezes, J., Prasad, U. & Ahmad, A. Elevated serum levels of transforming growth factor beta1 in Epstein-Barr virus-associated nasopharyngeal carcinoma patients. *Int. J. Cancer* **84**, 396–399 (1999).
26. Chou, J. *et al.* Nasopharyngeal carcinoma–review of the molecular mechanisms of tumorigenesis. *Head Neck* **30**, 946–963 (2008).
27. Sharpless, N.E. *et al.* Loss of p16Ink4a with retention of p19Arf predisposes mice to tumorigenesis. *Nature* **413**, 86–91 (2001).
28. Krimpenfort, P. *et al.* p15Ink4b is a critical tumour suppressor in the absence of p16Ink4a. *Nature* **448**, 943–946 (2007).
29. Lo, K.W. & Huang, D.P. Genetic and epigenetic changes in nasopharyngeal carcinoma. *Semin. Cancer Biol.* **12**, 451–462 (2002).
30. Wrensch, M. *et al.* Variants in the *CDKN2B* and *RTEL1* regions are associated with high-grade glioma susceptibility. *Nat. Genet.* **41**, 905–908 (2009).
31. Chen, M.C. *et al.* The incidence and risk of second primary cancers in patients with nasopharyngeal carcinoma: a population-based study in Taiwan over a 25-year period (1979–2003). *Ann. Oncol.* **19**, 1180–1186 (2008).
32. Scélo, G. *et al.* Second primary cancers in patients with nasopharyngeal carcinoma: a pooled analysis of 13 cancer registries. *Cancer Causes Control* **18**, 269–278 (2007).

ONLINE METHODS

Ethics. The study was approved by each of Institutional Review Board at the Sun Yat-sen University Cancer Center (SYSUCC) and the Genome Institute of Singapore. Informed consent was obtained from all study participants.

Study subjects. For the genome-wide analysis, 1,615 cases were recruited from southern China through the SYSUCC, which is located in the Guangdong province of southern China; this region is one of the endemic regions of NPC. All cases were histopathologically diagnosed by at least two pathologists according to the World Health Organization (WHO) classification. The cases are self-reported as Guangdong Chinese and lived in Guangdong province at the time of the study. All 1,025 healthy controls were recruited from local communities in the Guangdong province. All the controls were self-reported Guangdong Chinese and lived in Guangdong province at the time of the study and had no history of malignancy. The 1,008 Singapore Chinese controls were recruited as part of the Singapore Prospective Study Program; these individuals were self-declared Chinese and determined to be free of cancers through the use of questionnaires. Gender and age were collected from both the cases and the controls through questionnaires.

For the validation analysis, two independent case-control groups and one family trio group were recruited from the same Chinese population in southern China. The first group consisted of 2,800 cases and 2,100 controls of self-reported Chinese ancestry recruited in Guangdong province. The second sample consisted of 847 cases and 1,031 controls of self-reported Chinese ancestry recruited in Guangxi province (which is located next to Guangdong province). The third sample consisted of 284 family trios (probands and their parents) of self-reported Guangdong Chinese ancestry from Guangdong province. Recruitment details and sample inclusion criteria for each group are further described in the **Supplementary Note**.

Sample preparation and genotyping and quality control. Genomic DNAs were isolated from whole blood samples using a commercial DNA extraction kit (Qiagen) and quantified using PicoGreen reagent (Invitrogen). Genotyping analysis of the GWAS samples was conducted using Human610-Quad (for all the cases and the controls from Guangdong) and Human1M-Duo (for Singapore Chinese controls) BeadChips (Illumina).

Only the 576,998 autosomal SNPs common to both the Human610-Quad and Human1M-Duo BeadChips were analyzed. As a part of quality-control analysis, 22 samples with a SNP call rate of <96% were removed. 112,670 SNPs were excluded if they had a call rate < 95%, a minor allele frequency < 3% or significant deviation from Hardy-Weinberg Equilibrium in the controls ($P < 10^{-6}$). 464,328 SNPs in 3,626 samples passed quality control and were used for further analysis.

The genotyping analysis of the validation samples was done by using either TaqMan SNP genotyping assay (ABI) (for the three SNPs within the *MHC* region) or the MassArray system from Sequenom (for the non-*MHC* SNPs). The same SNP filtering criteria as the genome-wide analysis was applied, and all the individuals with SNP call rate less than 90% were removed from further analysis. In addition, for the 49 SNPs subjected to validation study, we examined the clustering patterns of genotypes from Infinium, TaqMan and Sequenom assays and confirmed the good quality of genotyping.

Quality control analysis. We examined potential genetic relatedness of the 3,626 GWAS samples using pairwise identity-by-state–based analysis in PLINK (v1.06)[33]. For each of the identified first- or second-degree relative pairs, the sample with the lower genotype call rate was removed. We further removed 33 individuals with sample heterozygosity more than three standard deviations from the mean of the percent of sample heterozygosity (**Supplementary Fig. 5**). Subsequently, we used PCA-based methods[34] to detect population outliers and stratification. As described previously[14], all the SNPs within five distinct regions of long-range LD were excluded from PCA analysis, including the *HLA* region on chromosome 6, inversions on chromosomes 8 and 5, and two regions on chromosome 11 (**Supplementary Table 16**). For the initial PCA analysis, all the 3,498 samples were analyzed together with the 206 reference samples or with 89 Asian samples from the International HapMap Project. Twenty-one population outliers were identified and removed (**Supplementary Fig. 1**). We then performed the second PCA analysis using the remaining 1,583 case and 1,894 control samples (**Supplementary Fig. 1**). Finally, 464,328 SNPs in 1,583 cases and 1,894 controls were used for genome-wide association analysis.

Statistical analysis. Genome-wide association analysis was performed by using the Cochran-Armitage trend test in a logistic regression model where age, gender and the first ten principal components from EIGENSTRAT[34] were included as covariates. The Manhattan plot of $-\log_{10} P$ was generated using Haploview (v4.1)[35]. The quantile-quantile plot was generated using R (see URLs) to evaluate the overall significance of the genome-wide associations and the potential impact of population stratification.

The Guangdong and Guangxi samples were both of southern Chinese origin, a group in which less population substructure was found[14,36]. For the validation analyses of the case-control samples (validation 1), the same trend test in a logistic regression model was first performed in the Guangdong and Guangxi groups using age and gender as covariates. In the joint analysis of the Guangdong and Guangxi samples, the study indicator (Guangdong or Guangxi) entered the model as an additional covariate. To combine the association evidence from the GWAS and the case-control validation samples, we treated the GWAS sample and the two independent validation case-control samples from Guangdong and Guangxi as independent studies and used the logistic regression model which included the covariates age, gender, study indicator (GWAS, Guangdong or Guangxi) and the first ten principal components from EIGENSTRAT[34] (the ten principal components were set to zero for the validation samples). The age and gender distributions for case-control samples are summarized in **Supplementary Table 17**.

For the family trios validation sample, TDT implemented in PLINK was used to evaluate differences between the transmitted and untransmitted allele counts in trios. To jointly analyze the case-control and trio samples, the log odds ratios (θ) and the corresponding variance (V) in the combined case-control (GWAS and replication 1) and family-based analyses were calculated using the inverse variance method[37]. The corresponding joint P values were calculated based on the Wald test, which assumes that the statistic of θ^2/V follows the χ^2 distribution[38].

The ORs calculated were the OR per allele and are presented for the minor allele of each SNP, unless it was stated otherwise elsewhere. Independence test of association was carried out in a conditional logistic regression analysis implemented in R.

Haplotypes were estimated using PHASE[39] in each study (the GWAS without Singapore Chinese control, the Guangdong case and control replication and the Guangxi case and control replication). The most likely haplotype pair was taken for each subject and the haplotype-based association analyses were carried out using logistic regression, which included haplotypes as variables and study indicator, age and gender as covariates.

Imputation. For the imputation of *HLA* class I alleles, we used the LD and haplotype information from the HapMap CHB samples to predict the *HLA-A*, *HLA-B* and *HLA-C* alleles among 1,583 cases and 972 controls of Guangdong Chinese ancestry in our GWAS samples using an approach similar to that used in previous studies[40,41]. In brief, we searched for a SNP combination that is in LD with the *HLA* allele of interest and inferred the haplotypes of these SNPs using the PHASE program[39,42] for case and control populations separately as suggested in a previous study[43]. The imputed *HLA* genotype, representing the estimated number of copies of the *HLA* allele, was calculated by summing the probability of having that allele given a specific haplotype, weighted by the corresponding haplotype probability[41].

For the imputation analysis in non-*MHC* loci, untyped genotypes were imputed in the GWAS samples by using IMPUTE (v1.0)[44] and the haplotype information from the HapMap CHB and JPT samples. The association test was performed using a logistic regression analysis as described above, where age, gender and ten principal components were used as covariates. Regional plots (**Fig. 2**) were generated using R to show the $-\log_{10} P$ and LD r^2 values with top SNP and recombination rates; the LD was calculated using PLINK; and the recombination rates (in cM/Mb unit) were estimated by using the inter-SNP genetic distance (cM) (downloaded from IMPUTE website and based on CHB+JPT population) divided by inter-SNP physical distance (Mb).

doi:10.1038/ng.601

Gene-gene interaction. The pairwise interaction analyses between the three confirmed *MHC* SNPs and the four confirmed non-*MHC* SNPs and among the four non-*MHC* SNPs were performed in the combined GWAS and validation case-control samples. The logistic regression model used for test in the interaction term also included age, gender and the first ten principal components from EIGENSTRAT as covariates (the ten principal components were set to zero for the validation samples).

33. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
34. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
35. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
36. Xu, S. *et al.* Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am. J. Hum. Genet.* **85**, 762–774 (2009).
37. Kazeem, G.R. & Farrall, M. Integrating case-control and TDT studies. *Ann. Hum. Genet.* **69**, 329–335 (2005).
38. Agresti, A. *Categorical Data Analysis* (John Wiley and Sons, New York, 1990).
39. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
40. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38**, 1166–1172 (2006).
41. Feng, B.J. *et al.* Multiple loci within the major histocompatibility complex confer risk of psoriasis. *PLoS Genet.* **5**, e1000606 (2009).
42. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
43. Mensah, F.K. *et al.* Haplotype uncertainty in association studies. *Genet. Epidemiol.* **31**, 348–357 (2007).
44. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).