Bioinformatics Analysis for Coding SNPs of the HLA-DQA1 Gene Involved in Susceptibility to Cervical Cancer

Yanyun Li
Jun Xing
Linsheng Zhao
Yanni Li
Yuchuan Wang
Weiming Zhang

Tianjin Medical University Basic Medical Research Center, Tianjin 300070, China.

Correspondence to: Yanyun Li
Tel: 86-22-235-2775
E-mail: springsunny@163.com

Received September 30, 2005; accepted November 25, 2005.

OBJECTIVE To analyze coding SNPs of the HLA-DQA1 gene involved in susceptibility for cervical cancer by a bioinformatics approach, and to choose some SNPs that may have an association with cervical cancer.
METHODS By a SNPper tool we extracted SNPs from a public database (dbSNP), exporting them in FASTA formats suitable for subsequent use. Then we used PARSESNP as a tool for the analysis of the cSNPs.
RESULTS In the cSNPs of the HLA-DQA1 gene, we find that rs9272693 and rs9272703, are made up of missense mutations which convert a codon for one amino acid into a codon for a different amino acid. We chose a PSSM Difference >10 as a lower level for the scores of changes predicted to be deldterious.
CONCLUSION We used a bioinformatics approach for cSNPs analysis of the HLA-DQA1 gene. This method can select the variants in a conserved region, and give a PSSM Difference score. But the results need to be verified in cervical cancer patients and a control population.

KEYWORDS: bioinformatics, single nucleotide polymorphisms, cervical cancer, HLA.

Cervical cancer is the third most common cancer in women worldwide.[1] Infection with oncogenic types of human papillomavirus (HPV) is the main cause of cervical cancer and its precursor lesions [cervical intraepithelial neoplasia (CIN)]. During their life-time many women become infected with HPV, but only a minority develop CIN or cervical cancer. Consequently, there have to be other factors, e.g., genetic factors, that play a role in the development of CIN or cervical cancer. Almost all research on cervical cancer susceptibility has focused on genes in the HLA-complex. The HLA-complex, on the short arm of chromosome 6 (6p21.3), contains Class-I, -II, -III and other 200 genes with known or unknown functions and a strong LD exists between them. The function of both Class-I and Class -II genes is the presentation of short, pathogen-derived peptides to T cells. The products of the Class-I genes (HLA-A, -B and -C) are usually associated with presentation of endogenous proteins. Class-II genes (HLA-DR, -DQ, and -DP) are associated with presentation of exogenous proteins.[2] Zoodsma et al.[3] identified all published studies from 1980 to January 2002 on the PubMed databases. They focused on common and genetic risk factors such as HLA and other genes (Tp53, IL-10, CYP2D6 and MTHFR) that may be involved in susceptibility to (pre) neoplastic cervical disease. We selected HLA-DQA1 for further analysis.

Single nucleotide polymorphisms (SNPs) are an increasingly im-

portant resource for understanding the structure and history of the human genome. A SNP is defined as a mutation involving a single DNA base substitution that is observed with a frequency of at least 1% in a given population. SNPs are the most common form of genetic variation in humans. Overall, SNPs account for 90% of inter-individual variability. [4] Scientific advancements have resulted in a series of genetic markers with ever-increasing information content and resolving power. In the past 30 years, restriction fragment length polymorphisms (RFLPs), short tandem repeats (STRs), and SNPs have played significant roles in genetic research. In order to analyze coding SNPs (cSNPs) in the HLA-DQA1 gene, which is a subtype of HLA-II, we planned to devise a platform to choose cSNPs by bioinformatics tools and predict variation that is likely to have a functional effect.

MATERIALS AND METHODS

Retrieval of human cSNPs for the HLA-DQA1 gene

SNPper is a web-based tool to automate the task of extracting SNPs from public databases, to analyze them and to export them in formats suitable for subsequent use.[5] SNPper is freely available at http://snpper.chip.org/. Registration is optional and it provides access to some advanced features. The most important public SNP database is dbSNP (accessible at http://www.ncbi.nih.gov/SNP/), that collects all SNPs detected by either computational methods (i.e. comparing matching sequences stored in databases like GenBank) or direct observation. The general purpose of SNPper is to create sets of SNPs responding to user-defined criteria. SNPs can be retrieved through their name. On the SNPper's interface we selected "Gene Finder", input gene's Symbol and then retrieved all the SNPs for the HLA-DQA1 gene. SNPper allows the user to filter or refine a SNP set to show the the cSNPs of the HLA-DQA1 gene. Lastly we can export the information that SNPper associates to each SNP.

Extracting FASTA sequences of human cSNPs for the HLA-DQA1 gene

FASTA sequences of human cSNPs for the HLA-DQA1 gene were retrieved from public database dbSNP(accessible at http://www.ncbi.nih.gov/SNP/). One can submit all the SNPrs# , then the web will automatically send the FASTA sequences to your E-mail within 24 h.

Searching for homology models of the HLA-DQA1 gene

In order to predict the severity of the effect of a missense change on function, we had to obtain the homology models. PARSESNP accepts blocks. Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. Information about Blocks is available at http://blocks.fhcrc.org/. We used Reverse-PSI BLAST Searcher as a tool to search for homology models of the HLA-DQA1 gene.[6]

Analyzing for cSNPs of the HLA-DQA1 gene

PARSESNP is a tool to display and analyze polymorphisms in genes and is available on the World Wide Web at http://www.proweb.org/parsesnp/. [7] Using a reference DNA sequence, an exon/intron position model and a list of polymorphisms, that information can be extracted from GenBank (http://ncbi.nlm.nih.gov). It determines the effects of these polymorphisms on the expressed gene product, as well as the changes in restriction enzyme recognition sites. The results were saved in our computer.

RESULTS

Results of human cSNPs for the HLA-DQA1 gene

Table 1 shows the SNP set export form of the HLA-DQA1 gene. It displays the SNP rs#, SNP position, band, distance from previous SNP, alleles, gene and role. From the column of distance from previous SNP, we can see that the density of SNPs is high at the chromosome 6p21.32 band, far from the average distance that SNPs occur at approximately every 1000 bases over the human population.

Result of homology models for the HLA -DQA1 gene

We looked for Blocks in the Blocks Database at the http://blocks.fhcrc.org/. The result is IPB001003,Class II histocompatibility antigen, alpha chain, alpha-1 domain, Score 382 bits and E Value e^{-107}.

Analysis results for cSNPs of the HLA-DQA1 gene

The results of PARSESNP can be viewed in a variety of different formats (Figs.1, 2, Table 2). Out of the 50 SNPs in the resulting database by SNPper, only 14 SNPs showed analytic results by PARSESNP, the other 36 SNPs showed no BLAST results on the genomic sequence or on the coding sequence. Fig.1 displays the locations of the polymorphisms in the gene (both coding sequence and genomic sequence). Table 2 de-

Table 1. The SNP set export form of the HLA–DQA1 gene

#SNP rs#	SNP position	Band	Distance from previous SNP	Alleles	Gene(s)	Role
rs9272430	chr6:32713235	6p21.32		A/C	HLA-DQA1	Coding exon
rs9469203	chr6:32713244	6p21.32	9	A/G	HLA-DQA1	Coding exon
rs9272431	chr6:32713249	6p21.32	5	C/T	HLA-DQA1	Coding exon
rs9272433	chr6:32713273	6p21.32	24	C/T	HLA-DQA1	Coding exon
rs9272688	chr6:32717075	6p21.32	3802	C/T	HLA-DQA1	Coding exon
rs9272689	chr6:32717083	6p21.32	8	A/G	HLA-DQA1	Coding exon
rs1071630	chr6:32717104	6p21.32	21	C/T	HLA-DQA1	Coding exon
rs1129753	chr6:32717108	6p21.32	4	C/T	HLA-DQA1	Coding exon
rs1048027	chr6:32717147	6p21.32	39	C/T	HLA-DQA1	Coding exon
rs9272691	chr6:32717170	6p21.32	23	A/G	HLA-DQA1	Coding exon
rs9272692	chr6:32717185	6p21.32	15	C/T	HLA-DQA1	Coding exon
rs9272693	chr6:32717190	6p21.32	5	C/T	HLA-DQA1	Coding exon
rs12722058	chr6:32717191	6p21.32	1	A/G	HLA-DQA1	Coding exon
rs9272694	chr6:32717192	6p21.32	1	G/T	HLA-DQA1	Coding exon
rs9272695	chr6:32717194	6p21.32	2	G/T	HLA-DQA1	Coding exon
rs9272696	chr6:32717200	6p21.32	6	A/T	HLA-DQA1	Coding exon
rs12722065	chr6:32717205	6p21.32	5	A/C/G	HLA-DQA1	Coding exon
rs9272699	chr6:32717207	6p21.32	2	A/C	HLA-DQA1	Coding exon
rs1048052	chr6:32717209	6p21.32	2	A/C/G/T	HLA-DQA1	Coding exon
rs12722070	chr6:32717211	6p21.32	2	C/T	HLA-DQA1	Coding exon
rs12722076	chr6:32717235	6p21.32	24	A/G/T	HLA-DQA1	Coding exon
rs9272706	chr6:32717249	6p21.32	14	C/G	HLA-DQA1	Coding exon
rs1048491	chr6:32717256	6p21.32	7	C/G/T	HLA-DQA1	Coding exon
rs12722081	chr6:32717259	6p21.32	3	A/C/G	HLA-DQA1	Coding exon
rs1048087	chr6:32717264	6p21.32	5	C/T	HLA-DQA1	Coding exon
rs1048124	chr6:32717761	6p21.32	497	C/T	HLA-DQA1	Coding exon
rs1048134	chr6:32717767	6p21.32	6	A/G	HLA-DQA1	Coding exon
rs707952	chr6:32717784	6p21.32	17	C/T	HLA-DQA1	Coding exon
rs9272745	chr6:32717784	6p21.32	0	C/T	HLA-DQA1	Coding exon
rs707951	chr6:32717791	6p21.32	7	C/T	HLA-DQA1	Coding exon
rs9272746	chr6:32717791	6p21.32	0	C/T	HLA-DQA1	Coding exon
rs1048173	chr6:32717833	6p21.32	42	A/C	HLA-DQA1	Coding exon
rs707950	chr6:32717851	6p21.32	18	C/G	HLA-DQA1	Coding exon
rs2308883	chr6:32717852	6p21.32	1	G/T	HLA-DQA1	Coding exon
rs707949	chr6:32717930	6p21.32	78	C/T	HLA-DQA1	Coding exon
rs2308885	chr6:32717942	6p21.32	12	G/T	HLA-DQA1	Coding exon
rs7990	chr6:32717943	6p21.32	1	A/C	HLA-DQA1	Coding exon
rs707963	chr6:32717947	6p21.32	4	G/T	HLA-DQA1	Coding exon
rs707962	chr6:32717952	6p21.32	5	G/T	HLA-DQA1	Coding exon
rs2308889	chr6:32717980	6p21.32	28	A/C	HLA-DQA1	Coding exon
rs2308890	chr6:32717986	6p21.32	6	C/T	HLA-DQA1	Coding exon
rs2308891	chr6:32717987	6p21.32	1	A/C/G	HLA-DQA1	Coding exon
rs9272785	chr6:32718379	6p21.32	392	A/G	HLA-DQA1	Coding exon
rs9272786	chr6:32718381	6p21.32	2	A/C	HLA-DQA1	Coding exon
rs9272787	chr6:32718414	6p21.32	33	C/T	HLA-DQA1	Coding exon
rs1048381	chr6:32718423	6p21.32	9	A/G	HLA-DQA1	Coding exon
rs1048414	chr6:32718456	6p21.32	33	C/G	HLA-DQA1	Coding exon
rs1048419	chr6:32718459	6p21.32	3	C/T	HLA-DQA1	Coding exon
rs1048430	chr6:32718465	6p21.32	6	C/G	HLA-DQA1	Coding exon
rs9272793	chr6:32718473	6p21.32	8	A/G	HLA-DQA1	Coding exon

Table 2. Table of polymorphisms for the HLA–DQA1 gene

Variant No.	Nucleotide Change	Effect	RE Gained in Variant	RE Lost from Reference	PSSM Difference	SIFT Score	Description	Zygosity
1	C3977T	R70W		AciI	20	0.01	rs9272693	Homo
2	A3987T	E73V		BseMII, DdeI	8.1	0.2	rs9272696	Homo
3	G3988T	E73D		BseMII, DdeI	2.5	0.51	rs9272697	Homo
4	A3995C	K76Q	CviJI	ApoI	-0.1	0.36	rs9272700	Homo
5	G4019T	G84C	BtsI,TspRI		15.1	0.12	rs9272703	Homo
6	G4020T	G84V	BspMI	ApaLI, BseSI, HgiAI, MjaIV, SduI	9.3	0.4	rs9272704	Homo
7	C4044T	A92V					rs9272709	Homo
8	C4051T	H94=					rs1048087	Homo
9	A4053C	N95T					rs1048089	Homo
10	A4058G	N97D					rs1048090	Homo
11	A4077C	Y103S	MmeI		0	0.99	rs9272711	Homo
12	C4571T	T130I	FokI		5.7	0.28	rs9272745	Homo
13	A5226G	M230V	BsiI	NlaIII			rs9272789	Homo
14	A5300C	P254=	ApaI, BseSI, DraII, HgiJII, NlaIV, SduI				rs9272794	Homo

Table of polymorphisms for the HLA-DQA1 gene shown in Table 2. The restriction enzyme names link to the entries in REBASE, describing commercial availability and isoschizomers. Descriptions are extracted from the dbSNP database. The last column shows the zygosity of the change, if a variant had been entered using an ambiguous nucleotide, this column would read 'hetero'.

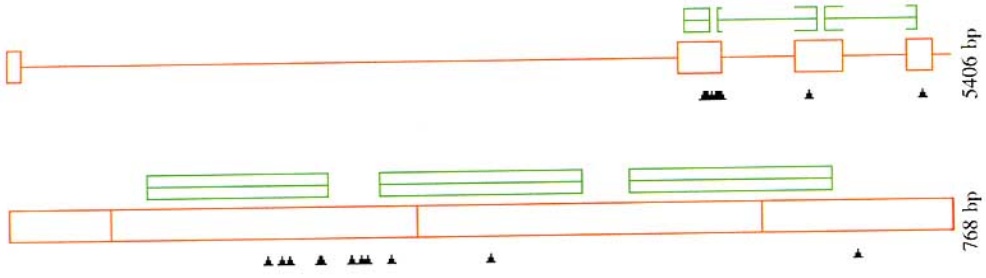

Fig.1. Overview of polymorphisms and Blocks in the HLA-DQA1 gene. (a) Genomic Sequence (b) Coding Sequence. Overview of polymorphisms and blocks in the HLA-DQA1 gene. The sequence and variants come from GenBank NC_000006.9 and the homology model derives from the Block IPB001003 families. In the 'Genomic Sequence' plot, the top region of the graphics shows the locations of the Blocks on the reference sequence. The green graphics correspond to blocks IPB001003A, IPB001003B and IPB001003C. The middle shows the locations of the exons, represented by boxes. The bottom region displays the locations of the polymorphisms, the first row displays missense changes in black and the second row shows silent changes in purple.

scribes in detail the effect of each polymorphism in the gene, including nucleotide change, amino acid result, restriction enzyme polymorphisms and PSSM difference. If the region containing a missense change is aligned to a block, one can attempt to gauge the effect of the change by examining the change in the PSSM score. We have chosen 10 as a rough lower level for the scores of changes predicted to be deleterious. Some nucleotide substitutions do not represent a change in the encoded amino acid, and are termed 'synonymous' cSNPs. Non-synonymous cSNPs can be those that result in conservative substitutions (amino acids with a similar size or charge) or non-conservative substitutions. Of the 14 SNPs, were 85.7% for

non-synonymous cSNPs, and 14.3% for synonymous cSNPs. In Table 2 rs9272693 and rs9272703 belong to non-synonymous cSNPs with a PSSM Difference >10 and SIFT Score <0.05. These have been empirically determined to be deleterious. Fig.2 displays the effect of the change form from the original amino acid, amino acid position, and new amino acid (*for stop codon). The variant of rs9272693 is the first nucleotide change C to T, which lead to the amino acid change from Arg to Trp. The variant of rs9272703 is the second nucleotide change G to T, which lead to the amino acid change of Gly to Cys.

DISCUSSION

The technical term, SNPs, appeared in human molecular genetic literature for the first time in 1994. SNPs are tightly associated with complex diseases. Association studies try to establish a relationship between a phenotype and one or more regions of the genome and the distribution and function of SNPs are important areas of current research. A variant may affect the expression or translation of a gene product, either by interrupting a regulatory region or by interfering with normal splicing and mRNA function. This can include SNPs in regulatory SNPs, intronic SNPs and exon-intron boundary SNPs. Research suggests that most SNPs fall in the 95 percent non-coding region of the genome with only 5 percent falling in the coding region. [8] Non-synonymous SNPs alter the amino acid substitution or introduction of a nonsense/truncation mutation.[9] The main purpose of this study was to predict the severity of the effect of a missense change on function. To assess the possible damaging effect of amino acid substitution, we developed a bioinformatics platform to analyze coding SNPs of the HLA-DQA1 gene involved in susceptibility for cervical cancer.

Today, the primary database of polymorphisms is dbSNP, which currently contains more than 5,000,000 validated human SNPs. A powerful resource for SNP analysis is SNPper. SNPper was created in the Kohane Lab at Harvard University for the analysis of SNPs. SNPper focuses on SNP selection for genetic studies and is freely available. Mooney[10] showed that general disease-associated mutations tend to occur in positions that are conserved. PARSESNP is a tool for the display and analysis of polymorphisms in genes. In order to assess the effect missense changes have on gene product function, PARSESNP provides a method of submitting homology information. PARSESNP accepts Blocks, a format that represents distinct regions of ungapped alignment in protein sequences from the Blocks database. The severity of the effect of a missense change on function can be predicted using the homology models.

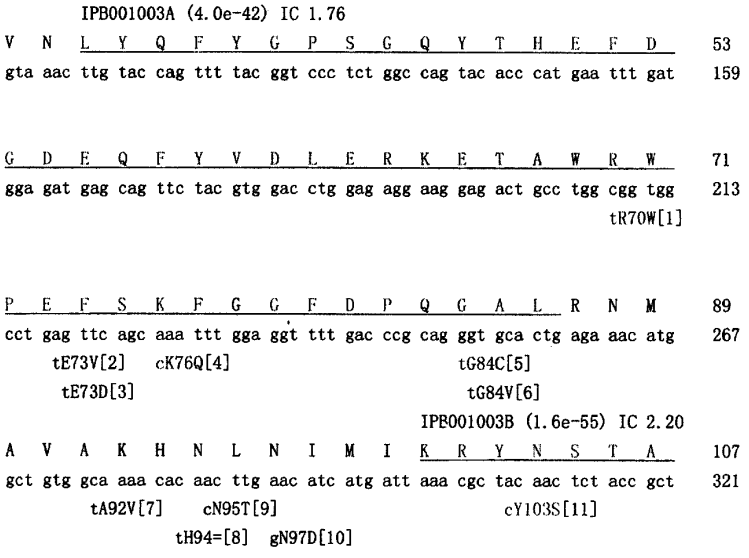

Fig.2. Polymorphisms in the sequence. Fig.2 Polymorphisms in the coding sequence from Fig.1 and Table 2. This region shows a Block aligned in the coding sequence. Two missense changes (number 1 and 5 in Table 2) in red and the other missense changes in black. Two missense changes (number 1 and 5) are colored red because the PSSM difference score is >10. The residues of the reference protein are colored to indicate how each position compares to the aligned Block. Those residues colored green are most similar to the corresponging column in the Block, while those colored red are most diverged.

There were 1446 SNPs in the HLA-DQA1 gene extracted by SNPper. The average distance was 17 nucleotides, which indicates that the density in this gene is very high. On coding sequence there are 50 SNPs, 14 SNPs of those produced analytic results by PARSESNP, another 36 SNPs could not get any BLAST results in the Genomic Sequence or coding sequence. A large number of redundant, incomplete, even incorrect SNPs are also collected in the SNP databases because of various resources of SNPs, such as: the results of sequencing, the BLAST of EST, variation in the results of experiments and so on. [11] Among the 14 SNPs which had analytic results, the PSSM Differences of rs9272693 and rs9272703 were more than 10. As for the prognosis of detrimental mutations, the probability of a deleterious variant is large if PSSM Difference >10.[7] The biologic significance is that the amino acid sequence coded by the variant nucleotide of rs9272693 and rs9272703 is changed, which probably alters the function of the HLA-DQA1 gene in production of exogenous proteins, and thus changes the immune reaction of an individual to HPV, and finally the susceptibility of cervical cancer increases.

In 2001, 1,420,000 SNPs in the human genome were reported in Nature by the International SNP Research Association and International Human Genome Sequencing Association. The data showed that SNPs occur approximately every 1250 bases. Up until the present, the number of known SNPs has grown at an ever increasing rate. After the Human Genome Project the study of SNPs has been a new focus of international research. Preliminary studies indicate that there are obvious differences between the Chinese and Western populations in the frequency of SNPs in several important diseases. Abundant hereditary resources in our country should be utilized to conduct SNPs research on important diseases with particular emphasis on constructing a genome SNPs' systematic catalogue of Chinese people. This would be quite meaningful for future health care and the biotechnologic medical industry. The number of SNPs is numerous and more than half are nonsense mutations,[10] so we need to increase the degree of success of experimental tests using bioinformatics tools to screen SNPs of people whose phenotypes and functions are altered. The solving of the problem depends on the development of the software and researchers' complete understanding and mastering of the software.

In this article we offer a set of practical, feasible approaches to solve the problem. The HLA-DQA1 gene studied in this report is based on previous research on cervical cancer. The mutations of the HLA-DQA1 gene may alter the function of the gene, and reduce the immune response of patients to HPV infection resulting in the promotion of cervical cancer. However this hypothesis needs further studies by measuring the frequencies of a number of SNPs in two populations, and by detecting SNPs that show a significant difference in frequency.

REFERENCES

1 Parkin DM, Pisani P and Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. Int J Cancer. 1999;80:827–841.
2 Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens. 2004;64:631–649.
3 Zoodsma M, Nolte I, Te Meerman GJ, et al. HLA genes and other candidate genes involved in susceptibility for (pre)neoplastic cervical disease. Int J Oncol. 2005;26:769–784.
4 Brookes AJ. The essence of SNPs. Gene. 1999;234:177–186.
5 Rica A, Kohane S. SNPper: retrieval and analysis of human SNPs. Bioinformatics. 2002;18:1681–1685.
6 Henikoff JG, Greene EA, Pietrokovski S, et al. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. 2000;28: 228–230.
7 Taylor NE, Greene EA. Parse SNP: a tool for the analysis of nucleotide polymorphisms. Nucl Acid Res. 2003;31: 3808–3811.
8 Xu L, Sun DZ, Yu ZH. Single nucleotide polymorphism of oncogenes and its individualized treatment. World Chin J Digestol. 2005;13:592–595.
9 Antonarakis SE, Cooper DN. Nature Encyclopedia of the Human Genome(Vol4). London:Nature Publishing Group. 2003:227–253.
10 Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief in Bioinform. 2005;6:44–56.
11 Wjst M. Target SNP selection in complex disease association studies. BMC Bioinformatics. 2004;5:92.