

DATABASE IN BRIEF

Thailand Mutation and Variation Database (ThaiMUT)

Uttapong Ruangrit^{1,7,†}, Metawee Srikumool^{2,†}, Anunchai Assawamakin^{1,3,†}, Chumpol Ngamphiw¹, Suparat Chuechote¹, Vilasinee Thaiprasarnsup², Gallissara Agavatpanitch¹, Ekawat Pasomsab⁴, Pa-thai Yenchitsomanus³, Surakameth Mahasirimongkol⁵, Wasun Chantratita⁴, Prasit Palittapongarnpim^{1,6}, Bunyarit Uyyanonvara⁷, Chanin Limwongse^{3,*}, and Sissades Tongsimma^{1,*}

¹*Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology (BIOTEC), Pathumtani, Thailand;* ²*Department of Biology, Faculty of Science, Chiang Mai University, Chiang Mai, Thailand;* ³*Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand;* ⁴*Department of Pathology, Faculty of Medicine, Ramathibodhi Hospital, Mahidol University, Bangkok, Thailand;* ⁵*Department of Medical Sciences, Ministry of Public Health, Center for International Cooperation, Foreign Affairs Section, Nontaburi, Thailand;* ⁶*Department of Microbiology, Faculty of Science Mahidol University, Bangkok, Thailand;* ⁷*Department of Information and Computer Technology, Sirindhorn International Institute of Technology (SIIT), Pathumtani, Thailand.*

† Equally contributing first authors

*Correspondence to Sissades Tongsimma, Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Pathumtani 12120, THAILAND
Tel.: +66-2-564-6700 ext. 5551, Fax: +66-2-564-6584; E-mail: sissades@biotec.or.th, siclw@mahidol.ac.th

Communicated by Alastair F. Brown

With the completion of the human genome project, novel sequencing and genotyping technologies had been utilized to detect mutations. Such mutations have continually been produced at exponential rate by researchers in various communities. Based on the population's mutation spectra, occurrences of Mendelian diseases are different across ethnic groups. A proportion of Mendelian diseases can be observed in some countries at higher rates than others. Recognizing the importance of mutation effects in Thailand, we established a National and Ethnic Mutation Database (NEMDB) for Thai people. This database, named Thailand Mutation and Variation database (ThaiMUT), offers a web-based access to genetic mutation and variation information in Thai population. This NEMDB initiative is an important informatics tool for both research and clinical purposes to retrieve and deposit human variation data. The mutation data cataloged in ThaiMUT database were derived from journal articles available in PubMed and local publications. In addition to collected mutation data, ThaiMUT also records genetic polymorphisms located in drug related genes. ThaiMUT could then provide useful information for clinical mutation screening services for Mendelian diseases and pharmacogenomic researches. ThaiMUT can be publicly accessed from <http://gi.biotec.or.th/thaimut> . © 2008 Wiley-Liss, Inc.

Received 18 July 2007; accepted revised manuscript 18 February 2008.

KEY WORDS: Thailand's NEMBD; database; SNP; genetic variation

INTRODUCTION

The existence of human mutation databases, such as Human Gene Mutation Database (HGMD; Cooper, et al., 1998; Krawczak, et al., 2000) and Locus-Specific Database (LSDB; Claustres, et al., 2002), has influenced the occurrence of regional initiatives in the discovering and cataloging of genetic mutations and variations. The importance of collecting mutations and variations, which affect genetic diseases in various ethnic communities, was also raised during a HUGO meeting as the Mutation Database Initiative (MDI; Cotton, 2000). Complex interplay among population history, mating patterns and natural selection resulted in an accumulation of population specific mutation which underlies many common genetic diseases in a particular population. Discovery of novel mutations in an ethnic-specific population and storing them are consequential to disease detection, mutation screening and genetic counseling. Different ethnics exhibit varying mutation spectrums. The varying spectrums could be the cause of the different occurrences of Mendelian diseases across populations. As an example, hemoglobin genes reveal various mutation spectrums in different populations and even within the same population (Old, et al., 2001; Sirichotiyakul, et al., 2003). It has been known that β -Thalassemia is highly prevalent in Southeast Asia and Mediterranean while Sickle Cell Anemia is prevalent in African descendents (Clark and Thein, 2004). Armed with the mutation spectrum information in a given population, researchers can devise an appropriate scheme for mutation detection in that particular population. Such mutation information will also be beneficial for carrier screening in Mendelian diseases. Consequently, practitioners and researchers will benefit from a well-maintained mutation database.

Ethnic specific mutation databases are necessary in the studies of comprehensive demographic history and patterns of migration flow (Patrinos, 2006). Comparison of available spectrums of genetic mutations across populations could facilitate genotype-phenotype correlation studies. Apart from the collection of ethnic specific mutations, public genotypic databases had revealed more than 10 million single nucleotide polymorphisms (SNPs) in human genome in different populations (Smigielski, et al., 2000). However, validating the existence of these SNPs in a certain population and capturing their frequencies are required to determine the population significance of these SNPs (Thorisson, et al., 2005). Upon completion of these tasks, researchers would be able to further investigate predisposing genetic factors influencing diseases in their specific population.

This system is regularly maintained and updated to promote the utilization of this database in the Thai human genetic society. It allows direct personal submission, which must go through a manual assessment by the Thai human genetic research community before publishing the results on ThaiMUT. This establishment was prepared to ensure that the data and web application services available on this website will be beneficial not only to the local research community but also research communities outside Thailand.

ETHNIC BACKGROUND

Thailand covers an area of 514,000 km² in the center of the Southeast Asian peninsula. It is bordered by Myanmar (Burma), Lao People's Democratic Republic (Laos), Cambodia and Malaysia, and has 2,420 km of coastline on the Gulf of Thailand and the Andaman Sea. Thailand stretches 1,650 kilometers from north to south, and 780 kilometers from east to west at its widest part (<http://www.un.or.th/thailand/geography.html>). The estimated population is 64 million, of which approximately 9.3 million live in Bangkok, the capital city, and its vicinity. The official language of the country is called "Thai" for which 94% of population use as their first language. Four major dialects of the Thai language are the dialects used in the central, northern, southern and northeastern regions. Northeastern dialect is closely related to the Lao language. In the four southern most provinces of Pattani, Satun, Yala and Naratiwat situated near the Malaysian border, majority of the population there is Muslim speaking "Pattani" Malay. In the mountainous area of the northern region, there are approximately 525,000 highland people or hill tribes who speak distinct languages. Ten to fifteen percent of populations have Chinese origin due to steady flows of immigration from China to Thailand during 1850 toward the end of the World War II. Thus, the Chinese population in Thailand was established as commerce and artisan communities throughout the country (<http://www.un.or.th/thailand/population.html>).

In terms of ethnic-specific Mendelian disease backgrounds, Thalassemia is the most common genetic disease in Thailand. Mutation carriers were estimated at 30–40% of the Thai population. The high prevalence of Thalassemia carriers results in more than 12,000 new cases of severely affected births annually in Thailand (Lagampan, et al., 2004). For other genetic related diseases, cancer is also a major health problem and has been the most common

cause of death since 1999 (Ministry of Public Health, 2004). In Thai men, liver cancer is the most common disease followed by lung cancer. On the other hand, cervix and breast cancer are the two top cancer types prevalent in Thai women (Sriplung, et al., 2003).

SYSTEM DESIGN AND IMPLEMENTATION

The design of ThaiMUT is based on a three-tier architecture model (client, application server and database). Figure 1 depicts the overall architectural design of ThaiMUT. On the client layer, we use PHP-based scripts as a CGI on the Apache web server to render the graphical web-interface. Mutations, SNPs and other human genome reference information are tabulated in a MySQL relational database. To simplify the SQL complexity, we created a web interface using CGI scripts in the application server to compose the queries on behalf of the users. Users will only need to enter a keyword such as gene name, disease name then the web server will transform the requests to query language (SQL) to get the pertinent data from MySQL.

The content and structure of ThaiMUT follows guidelines given by <http://www.hgvs.org> and by Scriver's recommendation (Scriver, et al., 1999). ThaiMUT was constructed in such a way that mutations/SNPs can be incorporated into and queried from the database. We began by incorporating large number of Thai mutation reports excerpted from literatures in PubMed to our MySQL database. Some of the mutations not in PubMed were collected from local publications and personal communication with researchers. In addition to mutations, validated SNPs in Thai population were also cataloged. These SNPs were obtained from (Mahasirimongkol, et al., 2006) which compared amongst Thais and Northeast Asian populations (Chinese and Japanese) their allele frequencies and linkage disequilibrium (LD) patterns from 188 drug related genes. This data set was obtained by genotyping SNPs 280 individuals from 4 major geographical regions in Thailand. Allele frequencies from these drug-related SNPs were systematically made available for the first time through the ThaiMUT database. Flanking sequences of these SNPs were compared against the latest build of reference sequence (RefSeq) obtained from NCBI database (Pruitt, et al., 2007). They can be visualized along with SNPs from other populations in various public domain databases namely dbSNP (Smigielski, et al., 2000), JSNP (Hirakawa, et al., 2002) and HapMap (Thorisson, et al., 2005).

Related genomics and genetic information are supplied along with the mutations and SNPs, for example: locus IDs, gene names, OMIM associated Mendelian diseases, nomenclatures, allele frequencies and their publication references with link-outs to PubMed. In order to assist scientists to compare SNPs across different populations, ThaiMUT integrated the latest public domain SNP databases. Therefore, SNP maps from each database can be presented all at once along with Thai SNPs. For graphical viewing, a W3 standard Scalable Vector Graphics (SVG; <http://www.adobe.com/svg>) was adopted; users can visualize the location and characteristic of selected SNPs from a comparative view of different populations. An apache web-server was set up to interact with the underlying database and to provide web-based graphical output to users. Users can query for both mutations and SNPs via regular web search box or intuitively select a locus from ideogram view to get the required information.

To regularly maintain data accuracy and to alert a curator to update the database, ThaiMUT features a direct submission of unpublished data from researchers. However, that data will be marked as unpublished and will be unmarked later when publication is officially made. The submission form complies with the guideline given by the Human Genome Variation Society (HGVS; <http://www.hgvs.org>). Most novel mutations and SNPs are expected to be submitted by members of the Thai Genetic Society who share common interests in Mendelian diseases and genetic epidemiological studies.

DATABASE ACCESS

ThaiMUT database is publicly available from <http://gi.biotech.or.th/thaimut>. Most popular web browsers, e.g., Internet Explorer (IE6 on Windows XP and IE7 on Vista) and Firefox, should be able to view mutation/variation contents in ThaiMUT. To access ThaiMUT, the JavaScript feature must be enabled. By default, web browsers enable JavaScript but disable other security-prone features such as ActiveX and pop-ups blocking. Since ThaiMUT relies only on JavaScript, most users should be able to access the web site. To graphically visualize SNPs on genes, Internet Explorer users are required to install SVG (Scalable Vector Graphic) viewer plug-in made available by Adobe software (<http://www.adobe.com/svg>).

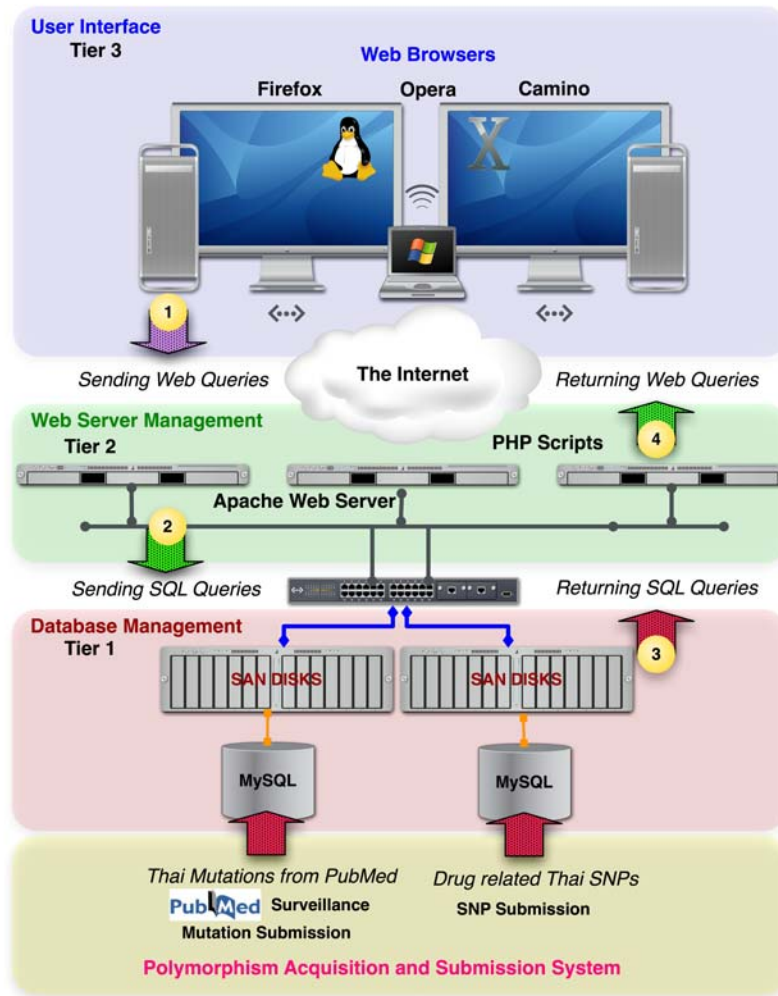


Figure 1. Implementation of ThaiMUT database system: 1) Users search for mutation/variation via web interface implemented using PHP scripts 2) Apache web server receives search requests and converts them to SQL commands issued to MySQL database 3) MySQL returns queried results back to Apache server 4) PHP scripts process raw outputs and transform them to html or SVG results that web browsers can render.

QUERYING THE DATABASE

For convenience and efficiency, both mutation and variation information in ThaiMUT must be queried through intuitive web interface similar to other genomic database interfaces. We have made the interface using Ext JS2.0 (extjs.com) which is a JavaScript library used in construction of many web applications. Four main feature search schemes (described below) are presented as tabs in the control frame of ThaiMUT. The database summary such as number, types of mutations is displayed in ThaiMUT welcome page. Figure 2 illustrates the ThaiMUT interface.

1. **Basic Search:** This is the most recommended feature search to explore mutations and variations in ThaiMUT. The search box can accept any string of gene name or locus ID, disease name, chromosome

number, OMIM number, nomenclature, title and author of reference article. The basic search is separately provided for mutation as well as variation.

2. **Advanced Search:** Similar to basic search, this feature accepts strings of all types. It also offers the search by mutation types and a range of years that an article was published. These queries can also be combined, e.g., user can search by gene name *and/or* author of the article at the same time.
3. **Alphabetical Search:** This feature arranges gene symbols or names in alphabetical order. Clicking on each alphabet, i.e., the initial of the gene name, will narrow the search space. All available genes with mutations found will be displayed.
4. **Chromosomal Search:** This feature offers a list of 22 autosomes plus the X and Y chromosomes. Users can directly jump to a chromosome of interest and visualize where mutations/variations are on the chromosome. Users can click on one of the colored boxes indicating where mutations/variations occurred on the chromosome.

Menu

- Mutation Basic Search (1)
- Variation Basic Search (2)
- Mutation Advanced Search (3)
- Mutation Alphabetical Search (4)
- Chromosomal Search (5)
- Submission Form (6)
- Help & Contact (6)

Biostatistics and Informatics Lab.,
Genome Institute,
National Center for Genetic Engineering
and Biotechnology (BIOTEC)
113 Thailand Science Park,
Phahonyothin Road, Klong 1, Klong
Luang, Pathumthani 12120 THAILAND
Tel: (66-2) 564 6700,
Fax: (66-2) 564 6701-5
email: uttapong.rua[at]biotec.or.th

Welcome

Reminder: This website requires JavaScript and SVG enabled browser to view its content. Please make sure you enable JavaScript before using the services. In addition, if you are not using either Firefox or Opera to view this page, please download and install SVG plug-in [here](#). To enable help function you have to unblock window popup and enable javascript for this site.

With the completion of the human genome project, novel sequencing and genotyping technologies had been utilized to detect mutations. Such mutations have continually been produced at exponential rate by researchers in various communities. Based on population's mutation spectrums, occurrences of Mendelian diseases are different across ethnic groups. A proportion of Mendelian diseases can be observed in some countries at higher rates than others. Recognizing importance of mutation effects in Thailand, we established a National and Ethnic Mutation Database (NEMDB) for Thai people. This database, named Thailand Mutation and Variation database (ThaiMUT) offers a web-based access to genetic mutation and variation information in Thai population. This NEMDB initiative is an important informatics tool for both research and clinical purposes to retrieve and deposit human variation data. The mutation data cataloged in ThaiMUT database were derived from journal articles available in PubMed and local publications. In addition to collected mutation data, ThaiMUT also records genetic polymorphisms located in drug related genes. ThaiMUT could then provide useful information for clinical mutation screening services for Mendelian diseases and pharmacogenomic researches.

Database Summary

Thailand Human Mutation Database Information	Thailand Human Variation Database Information
all mutations : 589	all SNPs : 1536
mutated genes : 119	number of genes containing SNPs : 228
phenotypes(diseases) : 119	SNPs grouped by function :
publications : 219	locus-region : 133
mutations grouped by type :	coding-synon : 54
substitution : 507	coding-nonsynon : 39
deletion : 59	mRNA-utr : 203
insertion : 20	intron : 1135
indels : 3	near-gene : 45
	(Sum of these SNPs > all SNPs due to multiple gene isoforms.)

© Biostatistics and Informatics Laboratory, Genome Institute, BIOTEC 2008. All Rights Reserved.

Figure 2. ThaiMUT web interface to the database: 1) Two tabs for basic search; one for mutation and the other for variation 2) Mutation advanced search 3) Search mutation by gene name ordered in alphabetical 4) Mutation search by chromosome, 5) Submission system and 6) Online help and contact.

DATA SUBMISSION

ThaiMUT encourages the human genetic research community both in Thailand and other countries to submit their discovered polymorphisms to the database. Users can click on the submission tab (see Fig. 2) to access the submission system. To submit either mutation(s) or variation(s), users are required to register their emails and contact information first. ThaiMUT will issue a password as in an alert box which can be copied and used to login to use the forms. The forms are provided separately for mutation and variation. In the design of ThaiMUT database, other types of mutations or variations such as STRs, microsatellites, minisatellites can be submitted and stored to the database. The submission form strictly follows proposed mutation entry and quality control form

(<http://www.hgvs.org/entry.html>), which is a recommendation by the MDI/HGVS. For security reasons, all submitted data would not be disclosed for public viewing unless they have been published or verified by a group of committees (e.g., appointed by the Thai Human Genetic Society), who are responsible for quality control of submission entries.

RESULTS AND DISCUSSION

Using the described design, the ThaiMUT database is now available for public access. There are now 119 causative genes identified in Thais registered in the database. A total of 589 mutations are listed, which can be categorized as 507 substitutions, 59 deletion, 20 insertions and 3 indels. Clearly, this is a much smaller collection than those in some other national and international mutation databases (George, et al., 2007). However, the ThaiMUT database is another example of a collaborative endeavor to gather and make available to the public the crucial resource regarding endemic genotyping information. This effort will eventually contain most of the identified causative mutation as well as SNP variants found in populations which are descendants of Thai ancestry. Currently, a number of causative mutations described for Thais can be readily searched for and used by this database. Moreover, attempts were made to expand the number of SNPs in the registry to maximize the opportunity for the future research in pharmacogenomics and disease association. Clinicians searching for information about a given genetic disease can easily browse through the ThaiMUT web interface to find gene information, mutation data, as well as information on local laboratories performing a particular genetic test. A comprehensive list of genes whose mutations are available in ThaiMUT is also displayed. This can indirectly serve as a national diagnostic laboratory service directory.

At present, data on 1312 SNPs from 188 drug-related gene loci have been deposited into the ThaiMUT database. Both common and rare alleles were included in this dataset. The study utilized tagSNPs, a minimum number of SNPs representing other SNPs in high linkage disequilibrium region, and common SNPs in these drug related gene loci to determine genetic distance between Thais and East Asian populations. Such tagSNPs were derived from the HapMap East Asian data set whose SNPs collocate with those of Thais'. Majority of these gene loci reveals similar haplotype spectrum and frequencies between Thai and East Asian populations suggesting that the tagSNPs could be informative for an indirect association study in these drug-related genes loci. However, for the minority of these loci that have different haplotype spectrum and frequencies, a small set of samples from a studied population should be used for SNPs discovery and tagSNPs selection before genotyping candidate polymorphisms for indirect association study. Underlying similarity information within these drug-related genes was deposited in the database for further utilization. Heterozygosity from multiple data sets in a certain population will be helpful in reducing the genotyping cost. In other words, we can avoid genotyping of uninformative (non-polymorphic) loci in the population. SNP discovery of a given genomic region is also important. Knowing the extent of population sequence variation diversity will be tremendously useful for the design of genetic association study (Tocharoentanaphol, et al., 2008).

With increasing amounts of genotype data in the future, the ThaiMUT database is certain to hold a central place for Thai genetic researchers for disease association as well as pharmacogenomics studies. Information contained in the mutation database can benefit both researchers and clinicians in the following ways. First, the data serves as a local reference for potential functionally significant variants that may be rare and have not been reported in other populations. This view justifies further functional studies of those particular variants rather than disregarding them as rare non-pathogenic polymorphisms. Second, researchers who are interested in a particular disorder can have an overview of spectrum of mutation found in Thailand as compared to those found elsewhere. This may give rise to a genetic mutation screening if a specific mutation is commonly found within our population. Third, using this database as an information portal, investigators having patients with similar genetic disorders or interested in similar genes can contact one another, share their resources and eventually undertake collaborative research projects. Moreover, this would potentially lead to a better cost-effective use of scarce resources for both physicians in Thailand and those neighboring countries interested in rare genetic diseases.

ACKNOWLEDGMENTS

The authors would like to thank all researchers who published mutations and variations that have been catalogued in ThaiMUT. We are also indebted to doctors from various hospitals in Thailand for sample collections as well as to their patients for their invaluable contributions. We would like to thank in advance the Thai Genetics Society who has kindly agreed to provide us with their mutation data in the future. We also thank the National Center for Genetic Engineering and Biotechnology (BIOTEC) for hosting the database and for supporting this research through the ThaiSNP Database Project (Grant No. BT-B-02-NT-BC-4728) and the Consolidation of BIOTEC Genome Database Project (Grant No. BT-B-02-IT-BC-4932). We also acknowledge the Thailand Research Fund (TRF) through the Royal Golden Jubilee Ph.D. Program (Grant No. PHD/4.I.MU.45/C.1) for supporting A.A. Finally we acknowledge the Thailand Center for Excellence in Life Sciences for funding support to S.M., E.P., and W.C.

REFERENCES

- Clark BE, Thein SL. 2004. Molecular diagnosis of haemoglobin disorders. *Clin Lab Haematol* 26(3):159-76.
- Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 12(5):680-8.
- Cooper DN, Ball EV, Krawczak M. 1998. The human gene mutation database. *Nucleic Acids Res* 26(1):285-7.
- Cotton RG. 2000. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. *Hum Mutat* 15(1):4-6.
- George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, Wouters MA, Cotton RG. 2007. General mutation databases: analysis and review. *J Med Genet* 45(2):65-70.
- Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. 2002. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res* 30(1):158-62.
- Krawczak M, Ball EV, Fenton I, Stenson PD, Abeyasinghe S, Thomas N, Cooper DN. 2000. Human gene mutation database a biomedical information and research resource. *Hum Mutat* 15(1):45-51.
- Lagampan S, Lapvongwittana P, Tharapan C, Nonthikorn J. 2004. Health Belief Model Teaching Program For Thalassemia Education in High School Students. *Chula Med Journal*.
- Mahasirimongkol S, Chantratita W, Promso S, Pasomsab E, Jinawath N, Jongjaroenprasert W, Lulitanond V, Krittayapoositpot P, Tongsima S, Sawanpanyalert P and others. 2006. Similarity of the allele frequency and linkage disequilibrium pattern of single nucleotide polymorphisms in drug-related gene loci between Thai and northern East Asian populations: implications for tagging SNP selection in Thais. *J Hum Genet* 51(10):896-904.
- Ministry of Public Health T. 2004. Public Health Statistics A.D. 2003.
- Old JM, Khan SN, Verma I, Fucharoen S, Kleanthous M, Ioannou P, Kotea N, Fisher C, Riazuddin S, Saxena R and others. 2001. A multi-center study in order to further define the molecular basis of beta-thalassemia in Thailand, Pakistan, Sri Lanka, Mauritius, Syria, and India, and to develop a simple molecular diagnostic strategy by amplification refractory mutation system-polymerase chain reaction. *Hemoglobin* 25(4):397-407.
- Patrinos GP. 2006. National and ethnic mutation databases: recording populations' genography. *Hum Mutat* 27(9):879-87.
- Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61-5.
- Scriven CR, Nowacki PM, Lehvaslaiho H. 1999. Guidelines and recommendations for content, structure, and deployment of mutation databases. *Hum Mutat* 13(5):344-50.

- Sirichotiyakul S, Saetung R, Sanguansermisri T. 2003. Analysis of beta-thalassemia mutations in northern Thailand using an automated fluorescence DNA sequencing technique. *Hemoglobin* 27(2):89-95.
- Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 28(1):352-5.
- Sriplung H, Sontipong S, Martin N. 2003. *Cancer in Thailand 3 (1995-1997)*(Bangkok Medical Publisher).
- Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The International HapMap Project Web site. *Genome Res* 15(11):1592-3.
- Tocharoentanaphol C, Promso S, Zelenika D, Lowhnoo T, Tongsimma S, Sura T, Chantratita W, Matsuda F, Mooney S, Sakuntabhai A. 2008. Evaluation of resequencing on number of tag SNPs of 13 atherosclerosis-related genes in Thai population. *J Hum Genet* 53(1):74-86.