Nucleic Acids Research, 2009, Vol. 37, No. Database Issue ***D1025–D1028***
doi:10.1093/nar/gkn966

The YH database: the first Asian diploid genome database

Guoqing Li[1], Lijia Ma[1,2], Chao Song[1], Zhentao Yang[1], Xiulan Wang[1], Hui Huang[1], Yingrui Li[1], Ruiqiang Li[1,3], Xiuqing Zhang[1], Huanming Yang[1], Jian Wang[1,*] and Jun Wang[1,3,*]

[1]Beijing Genomics Institute at Shenzhen 518083, [2]The Graduate University of Chinese Academy of Sciences, Beijing, 100062, China and [3]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark

Received August 12, 2008; Revised and Accepted November 14, 2008

ABSTRACT

The YH database is a server that allows the user to easily browse and download data from the first Asian diploid genome. The aim of this platform is to facilitate the study of this Asian genome and to enable improved organization and presentation large-scale personal genome data. Powered by GBrowse, we illustrate here the genome sequences, SNPs, and sequencing reads in the MapView. The relationships between phenotype and genotype can be searched by location, dbSNP ID, HGMD ID, gene symbol and disease name. A BLAST web service is also provided for the purpose of aligning query sequence against YH genome consensus. The YH database is currently one of the three personal genome database, organizing the original data and analysis results in a user-friendly interface, which is an endeavor to achieve fundamental goals for establishing personal medicine. The database is available at http://yh.genomics.org.cn.

INTRODUCTION

With the completion of the Human Genome Project, the life sciences, especially the field of genomics, has made major advances in the past several years. From the year of 2007, three personal genomes were sequenced and published as the pioneers of medical-oriented research (1–4). The sequence of the YH genome, the diploid Asian genome, was completed using only next generation sequencing technology—the Illumina Genome Analyzer (GA) (1). In comparison to the work carried out on the other two personal genomes, Craig Venter's and James Watson's genomes, the YH genome profited from using this new technology by being more cost effective and having higher output. With the aim to identify population-based polymorphisms that underlie complex diseases, the YH consensus genome was compared to the available reference human genomes (5,6) (the data of which are primarily of European descent), to detect SNPs and other structural variations.

In the database presented here: the YH genome and the processed variations were saved in standard format in order to enable data sharing and transformation, and to provide suitable utilities for analyses. The consensus genome sequence was saved as 'FASTA' with quality files and sequence variations were saved as 'gff3' files. This database is one of the first platforms created for personal genome data. Our efforts in designing the YH database were specifically geared to provide the best means for organizing and presenting personal genome data, and as such, is a useful resource for genomic and medical researchers.

DATA SOURCES AND METHODS

Data generation and SNP detection

The data placed in the database were generated using the Illumina GA, and resulted in 117.7 Gbp sequencing data of which 72 Gbp were single-end reads and 45.7 Gbp were paired-end reads. The genome was sequenced to a 36-fold average coverage (1). A total of 102.9 Gbp nucleotides were mapped onto the NCBI human reference genome (build 36.1) using the Short Oligonucleotide Alignment

*To whom correspondence should be addressed. Tel: +86 755 25273796; Fax: +86 755 25273796; Email: wangjian@genomics.org.cn
Correspondence may also be addressed to Jun Wang. Tel: +86 755 25273796; Fax: +86 755 25273796; Email: wangj@genomics.org.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Data summary of the YH genome sequence

Nucleotide	
Total	117.7 Gb
Map to genome	102.9 Gb
Coverage of genome[a]	99.97%
Polymorphism	
SNP	3.07 million
Indel	1 35 262
Structural variation	2682

[a]The fraction of reference genome which was covered by sequencing reads.

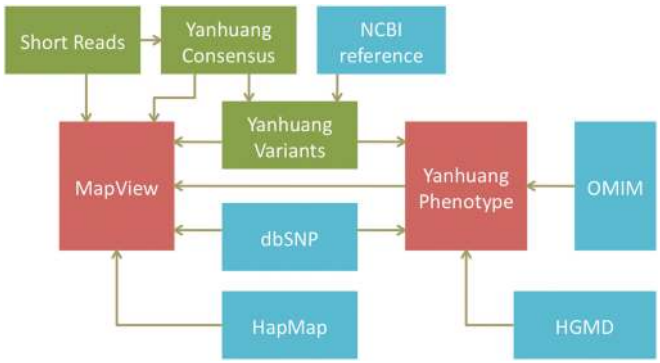

Figure 1. Data generation process.The sequencing reads were mapped to the NCBI reference genome to obtain the YH consensus genome and identify YH variants. Both types of data, combined with related information from dbSNP and HapMap, are presented in MapView. All variants were scanned by the disease alleles in HGMD to generate phenotypic information for the YH genome. Three colors of blocks were used to distinguish (i) data we generated (green); (ii) data from public database (blue); (iii) web pages we provided to show our genome (red).

Program (SOAP) (7). From this sequence, 3.07 millions SNPs were identified, and validated using Illumina 1M BeadChip (Table 1). The uniquely mapped reads, which were guided by the reference genome, assembled a high-quality consensus sequence for 92% of the Asian genome (92.6% of the autosomes, 83.1% of the sex chromosomes) (1). The data have been deposited in the EBI/NCBI Short Read Archive (Accession Number ERA000005) (1). SNPs and indels have been submitted to NCBI dbSNP and will be available in dbSNP version 130 (1).

Scanning for disease-related mutations

All YH SNPs were scanned with the academic version of HGMD (8), one of the best available mutation databases, which contains 53 643 variations and their corresponding phenotypes. HGMD uses internal IDs to identify mutations, we therefore aligned the mutations with flanking sequences to the reference genome to get their exact positions. We confirmed them with dbSNP identifiers. From this we identified a total of 1495 SNPs within 116 genes that are known to be associated with increased risk to tobacco addition and several diseases (1) (Figure 1).

DATABASE CONTENT AND ORGANIZATION

Consensus genome and SNPs

In the YH Database, the consensus genome sequence and the 3.07 million identified SNPs can be displayed in MapView, which was designed based on GBrowse (9) developed by GMOD (Figure 2). The visualization window was set at the default of 80 bp (depends on the used OS and web browser). To facilitate medically oriented studies, we integrated OMIM mutations, SNP frequencies of HapMap population, and YH genotypes in the MapView, using dbSNP ID as a cross-reference (10,11). The SNP frequency identified in the HapMap Project was displayed in a pie chart to enable users to easily check whether the YH genotype is consistent with the CHB population. Because deep sequencing was used, the visualization window shows all sequencing reads currently mapped to present alignments and highlights the unmatched nucleotides in each read. Around 190 Gbp reads were imported to our database, and we modified the EST module of GBrowse to implement a real-time presentation of reads alignments.

Mutations and phenotypes

In the total 53 643 HGMD identifiers were used to screen YH SNPs in order to retrieve phenotype related information. Phenotypes/diseases were categorized into 18 classes according to their physical features. Disease names, record numbers, and additional detailed information are displayed in a tree style. For each disease, we list related records in a following frame, including genomic position, dbSNP ID, HGMD ID, gene symbol, reference allele, risk allele and YH allele. As most records in HMGD have been only identified using the internal HMGD IDs rather than the dbSNP IDs, it is recommended to examine mutations of interest in the MapView to obtain detailed genomic information, which can be done by clicking on their chromosomal position.

Web services

We provide BLAST online to allow users to align customized query sequences to the YH consensus sequences. The sequences can be prepared in a FASTA format and pasted in query window or uploaded from a local disk.

Interface and access

The YH database is developed and maintained by BGI-shenzhen, which is a non-profit academic institution and can be accessed freely by the public. MySQL and JSP were used in database construction and interface utility, respectively.

A search can be done by inserting a genomic location, a gene symbol, a dbSNP ID or a YHSNP ID. In MapView, the primary display window, all basic information about the YH genome is shown in a genome browser. This browser contains Entrez genes, OMIM associated diseases, SNPs in dbSNPs, SNPs in HapMap, and YH genotypes. All sequencing reads presented in this window are set up in a manner to facilitate the users for checking each detected SNPs. The reads are overlapped one by one, and

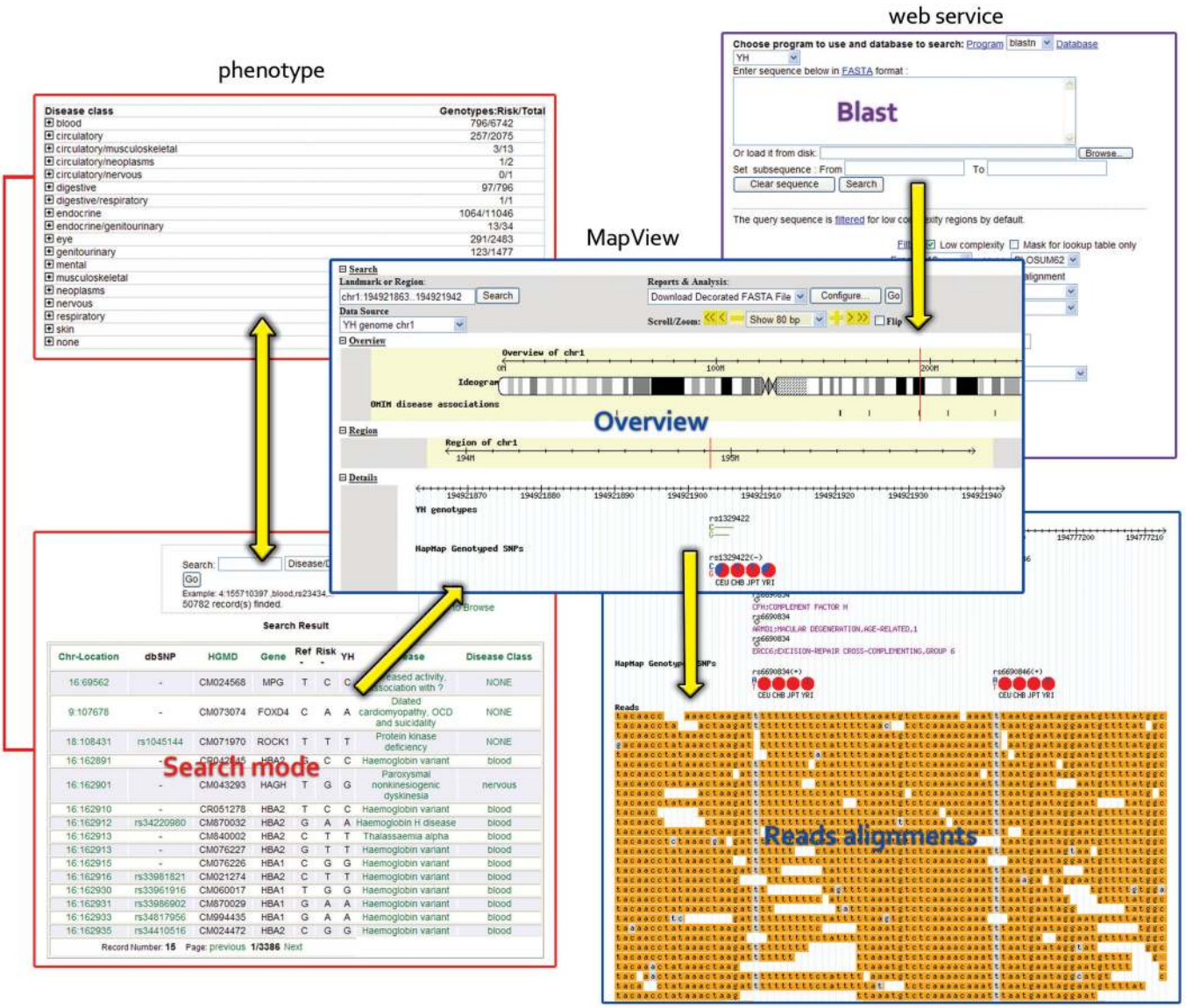

Figure 2. YH database Screenshot. We provide MapView, phenotype, and BLAST web service as the three main functions of the YH database. Users can easily begin searching or browsing this database for a variety of things such as a phenotype of interest or a risk allele. Access to the MapView can be obtained by clicking its chromosomal location. A BLAST web service for aligning query sequences against the YH consensus sequence is also available. All detailed information of the YH consensus sequence and identified SNPs are presented in the MapView. All sequencing reads mapped to an area in the display window are shown and are overlapped individually to facilitate users in examining each detected SNPs.

the unmatched nucleotides are highlighted in grey. Customized annotation files can be uploaded in a text format or from a remote URL. The browser window can be easily justified by changing the image width and region size, and by highlighting regions of interest.

On the phenotype page, a search view and a browse view option are both provided. All diseases are categorized into 18 classes, depending on their physical functions. When a user inserts a key word(s), detailed results are listed in a table that can be re-ordered by chromosome location, gene symbol, and disease name.

The BLAST services in the database that are provided for alignment include most of the main Blast programs (BlastN, BlastX, TBlastN and TBlastX) (12). Additionally it is easy to obtain help about BLAST usage by clicking on available external links.

The database contains all raw and processed data, including YH genome sequence, YH variants, annotations and short reads alignments, and can all be downloaded. The genome sequence and alignments within the database are also presented in a well-organized fashion by chromosome.

FUTURE WORK

In addition to the SNPs and SNP-related phenotypes have been shown in the YH database, there are 1 35 262 indels and 2682 structural variations, which are newly identified

and have not been presented in the MapView currently. However, these data have been available for downloading. Our group is working on developing new modules to display indels and kinds of structural variations in a visualized and friendly way, as the existing modules of GBrowse cannot implement it. Moreover, the ongoing project, Yanhuang 99, is generating more individual Asian genomes data, which will be imported to the YH database in the future. And more comparative analysis and population genetics studies will be put into practice and the results will be available in YH database also.

ACKNOWLEDGEMENTS

This work was supported by the Shenzhen Municipal Government and the Yantian District local government of Shenzhen.

FUNDING

National Natural Science Foundation of China (30725008; 90608010; 30221004; 90612019); the Chinese 973 program (2007CB815701; 2007CB815703; 2007CB815705); the Chinese 863 program (2006AA02Z177; 2006AA02Z334; 2006AA10A121); the Beijing Municipal Science and Technology Commission (D07030200740000); Ministry of Education (XXBKYHT2006001); the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council, the pig bioinformatics grant from Danish Research Council and the Solexa project (272-07-0196). Funding for open access charge: 2006AA02Z177.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
2. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
3. Brenner,S.E. (2007) Common sense for our genomes. *Nature*, **449**, 783–784.
4. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
5. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
7. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
8. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
9. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
10. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
11. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.