BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 6 2004, pages 989–992
DOI: 10.1093/bioinformatics/bth051

AmpliBASE *MT™: a* Mycobacterium tuberculosis *diversity knowledgebase*

Ahmed A. Majeed[1], Niyaz Ahmed[1], K. Rajender Rao[1], S. Ghousunnissa[1], Farhana Kauser[1], Buddhaditta Bose[1], H. A. Nagarajaram[1], V. M. Katoch[2], Debby V. Cousins[3], Leonardo A. Sechi[4], Robert H. Gilman[5] and Seyed E. Hasnain[1,6,]*

[1]Centre for DNA Fingerprinting and Diagnostics, Nacharam, Hyderabad 500076, India, [2]National Mycobacterial Repository Centre, Central JALMA Institute for Leprosy (ICMR), Agra 282001, India, [3]Australian Reference Laboratory for Bovine Tuberculosis, Agriculture Western Australia, South Perth, Australia, [4]Department of Biomedical Sciences, University of Sassari, Sardinia, Italy, [5]The Johns Hopkins Bloomberg School of Public Health and Hygiene, Baltimore, USA and [6]Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore 560064, India

Received on January 2, 2003; revised and accepted on November 12, 2003
Advance Access publication February 5, 2004

ABSTRACT

Summary: *AmpliBASE* MT™ is an online databank of high-resolution DNA fingerprints representing fluorescent amplified fragment length polymorphism (FAFLP) profiles or *amplitypes* developed for the *Mycobacterium tuberculosis* complex strains from 48 different countries. *AmpliBASE* MT™ is based on a relational database management system that is hyperlinked to visualize genotyping results in the form of DNA fingerprint images for individual strains. A flexible search system based on systematic comparisons of fragment sizes in base pairs allows inter-laboratory comparison of FAFLP profiles. Besides this, the database also displays previously published data on IS6110 profiles, spoligotypes, MIRU-VNTRs and large sequence polymorphisms along with the FAFLP records that will give the overall comparisons. Being the first of its kind, *AmpliBASE* MT™ is expected to be a very helpful tool in strengthening the concept of 'geographic genomics' and will be very helpful to molecular epidemiologists and those interested in diagnostic development for tuberculosis.

Availability: http://210.212.212.4/index.html

Contact: amplibas@cdfd.org.in

INTRODUCTION

DNA fingerprinting technology has emerged as a new generation approach for molecular typing of pathogenic bacteria. Restriction fragment length polymorphism (RFLP) analysis using IS6110 (Thierry *et al*., 1990) based probes has been recognized as a gold standard for typing *Mycobacterium tuberculosis* clinical isolates. This technique, however, has several disadvantages, primarily its inability to be used for isolates that are negative for the presence of IS6110, or carry only a single copy of the IS sequence (Das *et al*., 1995; Siddiqi *et al*., 2001). Interpretation of results is often difficult with IS6110 typing when the copy numbers of insertion fragment are fewer than 6 (Kremer *et al*., 1999; Siddiqi *et al*., 2001). While several other genome composition based techniques have been reported, there have been concerns regarding problems of portability, reproducibility and inter-laboratory comparisons of data. Recently, the technique of fluorescent amplified fragment length polymorphism (FAFLP) typing has been described for *M.tuberculosis* genotyping and molecular epidemiology (Goulding *et al*., 2000; Sims *et al*., 2002; Ahmed *et al*., 2003a). This technique has been very successfully used in typing various outbreak-associated strains (Ahmed *et al*., 2003a,b) and also has been used for describing new species at the molecular level (Ahmed *et al*., 2003b). Due to globalization of the TB epidemic it has become necessary to establish multinational and multicentric collaborations, where a large number of isolates are typed and compared between various laboratories. We now describe the development of a novel, multicentric, collaborative database system based on genome wide analysis of *Eco*RI/*Mse*I microrestriction patterns or FAFLPs in clinical isolates of the *M.tuberculosis* complex. We have named it *AmpliBASE* MT™ (trademark protected). The database makes it convenient for the interested scientific community to retrieve and deposit FAFLP based genotypic signatures of the isolates for outbreak investigation, molecular epidemiology and evolutionary studies involving high-resolution molecular markers. Besides this, *AmpliBASE* MT™ also allows corroboration of FAFLP data with IS6110 copy number change, spoligotype

*To whom correspondence should be addressed.

Downloaded from bioinformatics.oxfordjournals.org by guest on October 11, 2010

patterns (Kamerbeek *et al*., 1997) polymorphisms due to MIRU-VNTR elements (Supply *et al*., 2001) and large genomic deletions (Brosch *et al*., 2002). The database has been designed keeping in view the nature of complexity of upstream experiments involving genotypic analyses of hundreds of clinical isolates and is likely to be increasingly populated with additional FAFLP data and upgraded continuously in the future. The first release of *AmpliBASE* MT™ has been hosted via the Centre for DNA Fingerprinting and Diagnostics (CDFD) Web server (The EMBnet India node) and is freely accessible to TB control officers, scientists, students and technicians working in health departments, reference laboratories, academia and non-profit making organizations.

DATABASE DEVELOPMENT AND IMPLEMENTATION

The design of the database is illustrated in Figure 1. DNA fingerprints of *M.tuberculosis* strains developed by using the patented AFLP technology (Vos *et al*., 1995) were archived in the form of a searchable database.

Development of high-resolution fingerprint profiles

All the fingerprints were developed as described earlier (Ahmed *et al*., 2003a,b; Goulding *et al*., 2000) using the ABI Prism 377 XL DNA sequencer (Applied Biosystems) or the ABI Prism 3100 DNA sequencer (Applied Biosystems). DNA fingerprinting data were processed by Genescan Analysis® software (Applied Biosystems) with the help of the internal size ladder, GS ROX-500 (Applied Biosystems). Genotypic comparisons were performed by the allele calling software, Genotyper® (Applied Biosystems). Strain specific FAFLP patterns as seen in a Genotyper® plot window were captured as jpeg images and were then uploaded onto the Windows NT server through the Apache Tomcat as a JAVA servlets container.

Database design

A MySQL database based on relational database management system (RDBMS) was developed to archive FAFLP profiles along with other useful data relevant to the profiles. All the supporting data related to the DNA fingerprint profiles, such as strain name, species type, date of isolation, IS6110 profiles, MIRU-VNTR profiles and large sequence polymorphism status were arranged in the MySQL (DuBois, 2000) database under a master table. Four separate tables in MySQL were developed to contain path information for the images directory and for the text data related to fragment sizes in base pairs for all the FAFLP fingerprint profiles corresponding to four different primer sets used (*Eco*RI + A/*Mse*I+0, *Eco*RI+ G/*Mse*I+0, *Eco*RI+C/*Mse*I + 0, *Eco*RI + T/*Mse*I + 0). The database was queried from an html client using a JAVA servlets application as middle-ware for a platform independent, quick display of the records as dynamically generated web pages in different frames. Records were positioned, edited, retrieved and displayed with various algorithms that were written in C language. The search engine was developed to dynamically display results of searches based on size similarity of the FAFLP fragments using input data as molecular weight values in base pairs. The search engine was also made flexible to accept queries based on IS6110 copy numbers. The query processing was developed to handle multiple queries running on the database simultaneously.

The user interface

The *AmpliBASE* MT™ user interface (Fig. 1B) provides user friendly menus and frames to display genotypic images when the database is queried with country names or searched for similarities using base pair values (alphanumeric) as input data for various strains. Frames in Web browsers are adjusted to load the genotypic profile of the standard reference strains such as *M.tuberculosis* H37Rv, in the upper panel of the interface to allow visual inspection and comparison with genotypic patterns of different field strains that can be loaded in the lower panel.

DISCUSSION

We have described for the first time a searchable database of high-resolution FAFLP patterns developed for the *M.tuberculosis* complex. The strategy was developed under the auspices of the *M.tuberculosis* evolutionary genomics interest group at CDFD for facilitating easy and fast access to the different genotypic data generated in collaborating laboratories. *AmpliBASE* MT™ is fully capable of providing a large number of comparisons for *M.tuberculosis* genotypes to foster comprehensive evaluation of different strain typing methods employed for the identification and grouping of patient isolates. The portal will be able to provide better understanding of genome derived markers in terms of transmission, evolution and pathogenesis.

Although *AmpliBASE* MT™ is still in its growth phase, it promises quite significant benefits for molecular epidemiologists and disease control authorities in identifying prevalent, emerging and re-emerging strains of *M.tuberculosis* in various parts of the world. It is expected that users of *AmpliBASE* MT™ and collaborating laboratories will be able to deposit their genotypic data for reference purposes and the database will be populated on a routine basis. At present, *AmpliBASE* MT™ stores data on *Eco*RI/*Mse*I FAFLP patterns, for which commercial kits and reagents are easily available and the protocol is in widespread use in various laboratories. However, it can be revised in future to include many other restriction enzyme combinations to increase genome coverage.

The search engine of the database has been made flexible to display results differing by 1.0 bp keeping in view the sizing differences arising on different genetic analyzers (gel based versus capillary based systems) used in different laboratories.

Downloaded from bioinformatics.oxfordjournals.org by guest on October 11, 2010

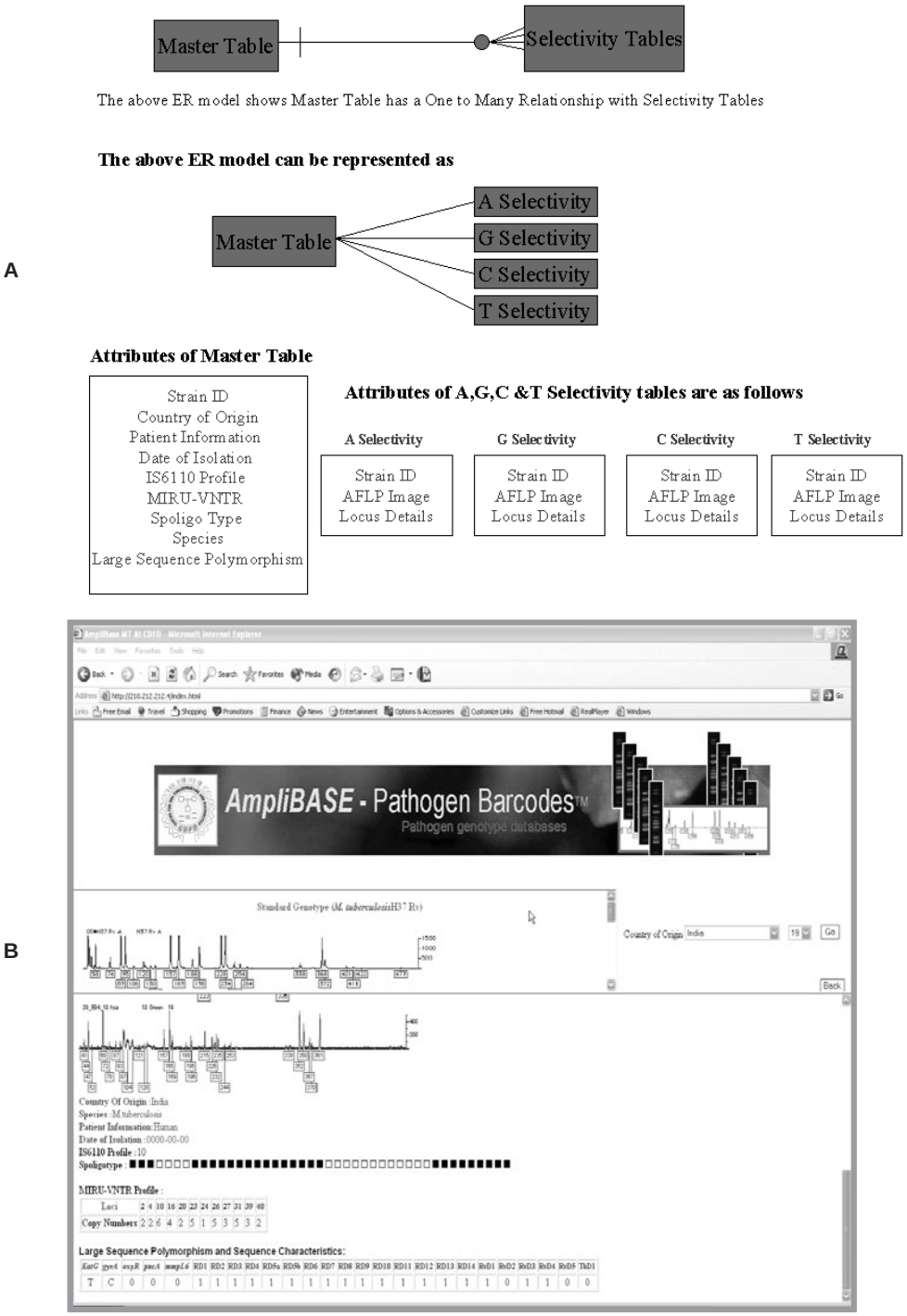

Fig. 1. **(A)** Schematic design of *AmpliBASE* MT™ database server. FAFLP data corresponding to four selective primers used are stored under separate tables named A selectivity, G selectivity, C selectivity and T selectivity. The master table housed all the clinical and microbiological details, comparative genotypic data and path information for FAFLP images. Information available from other typing systems such as spoligotyping, IS6110 RFLP, MIRU-VNTR analysis and large sequence polymorphisms also comprised the contents of the master table. Strain ID is the principal key for location of the records, deposition and editing. Along with every image belonging to either of the four categories, complete details of the genomic locus for each of the fragments in the fingerprint were also archived in the sub-tables. All the four sub-tables were accessible through the master table to arrange all the four images representing four individual fingerprints along with related records, according to the query and to display them as dynamically generated Web pages. **(B)** Screen dump of *AmpliBASE* MT™ user interface.

Downloaded from bioinformatics.oxfordjournals.org by guest on October 11, 2010

A large number of genotypes from a very impressive reference collection of strains representing Beijing, Haarlem and African clusters (Kremer *et al*., 1999) have been housed in the database to allow extensive comparisons.

In conclusion, *AmpliBASE* MT™ promises to be a very efficient online tool for mycobacteriologists across the world. It harnesses the already accepted and much appreciated technology of AFLP in the context of the *M.tuberculosis* genome. Future releases are likely to enhance the quality and speed of searches on the portal. The concept and ontology of *AmpliBASE* MT™ will be helpful for future efforts targeting other pathogenic microorganisms of major public health significance.

ACKNOWLEDGEMENTS

Financial support from WHO/TDR, International Society for Infectious Diseases (ISID), and core grants to CDFD from the Department of Biotechnology, Government of India is acknowledged. We thank D. van Soolingen and Kristin Kremer for providing a collection of DNA samples from internationally accepted reference strains. Support received from CDFD Bioinformatics team (Ranjit Kumar Gundu, Phani Prasad and Prashanthi) in implementation and hosting the database is acknowledged with thanks. We thank Yogesh Prasad (Applied Biosystems/Labindia) for helpful suggestions and discussions.

REFERENCES

Ahmed,N., Caviedes,L., Alam,M., Rao,K.R., Sangal,V., Sheen,P., Gilman,R.H. and Hasnain,S.E. (2003a) Distinctiveness of *Mycobacterium tuberculosis* genotypes from human immunodeficiency virus type 1-seropositive and -seronegative patients in Lima, Peru. *J. Clin. Microbiol.*, **41**, 1712–1716.

Ahmed,N., Alam,M., Majeed,A.A., Rahman,S.A., Cataldi,A., Cousins,D. and Hasnain,S.E. (2003b) Genome sequence based, comparative analysis of the fluorescent amplified fragment length polymorphisms (FAFLP) of tubercle bacilli from seals provides molecular evidence for a new species within the *Mycobacterium tuberculosis* complex. *Infect. Genet. Evol.*, **2**, 193–199.

Brosch,R., Gordon,S.V., Marmiesse,M., Brodin,P., Buchrieser,C., Eiglmeier,K., Garnier,T., Gutierrez,C., Hewinson,G., Kremer,K. *et al.* (2002) A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl Acad. Sci., USA*, **99**, 3684–3689.

Das,S., Paramasivan,C.N., Lowrie,D.B., Prabhakar,R. and Narayanan,P.R. (1995). IS6110 restriction fragment length polymorphism typing of clinical isolates of *Mycobacterium tuberculosis* from patients with pulmonary tuberculosis in Madras, South India. *Tuber. Lung Dis*., **76**, 550–754.

DuBois (2000) *MySQL*. New Riders, Indianapolis, IN, USA.

Goulding,J.N., Stanley,J., Saunders,N. and Arnold,C. (2000) Genome-sequence-based fluorescent amplified-fragment length polymorphism analysis of *Mycobacterium tuberculosis*. *J. Clin. Microbiol*., **38**,1121–1126.

Kamerbeek,J., Schouls,L., Kolk,A., van Agterveld,M., van Soolingen,D., Kuijper,S., Bunschoten,A., Molhuizen,H., Shaw,R., Goyal,M. and van Embden, J. (1997) Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol*., **35**, 907–914.

Kremer,K., van Soolingen,D., Frothingham,R., Haas,W.H., Hermans,P., Martín,C., Palittapongarnpim,P., Plikaytis,B., Riley,L., Yakrus,M. A., Musser,J.M. and van Embden,J.D.A. (1999) Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains. Interlaboratory study of discriminatory power and reproducibility. *J. Clin. Microbiol*., **37**, 2607–2618.

Siddiqi,N., Shamim,M., Amin,A., Chauhan,D.S., Das,R., Srivastava,K., Singh,D., Sharma,V.D., Katoch,V.M., Sharma,S.K., Hanief,M. and Hasnain,S.E. (2001) Typing of drug resistant isolates of *Mycobacterium tuberculosis* from India using the IS6110 element reveals substantive polymorphsim. *Infect. Genet. Evol*., **1**, 109–113.

Sims,E.J., Goyal,M. and Arnold,C. (2002) Experimental versus *in silico* fluorescent amplified fragment length polymorphism analysis of *Mycobacterium tuberculosis*: improved typing with an extended fragment range. *J. Clin. Microbiol.*, **40**, 4072–4076.

Supply,P., Lesjean,S., Savine,E., Kremer,K., van Soolingen,D. and Locht, C. (2001) Automated high-throughput genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.*, **39**, 3563–3571.

Thierry,D., Cave,M.D., Eisenach,K.D., Crawford,J.T., Bates,J.H., Gicquel,B. and Guesdon,J.L. (1990) IS6110, an IS-like element of *Mycobacterium tuberculosis* complex. *Nucleic Acids Res*., **18**, 188.

Vos,P., Hogers,R., Bleeker,M., Reijans,M., van de Lee,T., Hornes,M., Frijters,A., Pot,J., Peleman,J., Kuiper,M. and Zabeau,M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res*., **23**, 4407–4414.

Downloaded from bioinformatics.oxfordjournals.org by guest on October 11, 2010