Databases

Italian Rett Database and Biobank

Katia Sampieri,1 Ilaria Meloni,1 Elisa Scala,1 Francesca Ariani,1 Rossella Caselli,1 Chiara Pescucci,1 Ilaria Longo,1 Rosangela Artuso,1 Mirella Bruttini,1 Maria Antonietta Mencarelli,1 Caterina Speciale,1 Vincenza Causarano,1 Giuseppe Hayek,2 Michele Zappella,2 Alessandra Renieri,1,* and Francesca Mari1

1Medical Genetics Unit, Department of Molecular Biology, University of Siena, Siena, Italy; 2Child Neuropsychiatry, Azienda Ospedaliera Senese, Siena, Italy

Communicated by Richard Cotton

Rett syndrome is the second most common cause of severe mental retardation in females, with an incidence of approximately 1 out of 10,000 live female births. In addition to the classic form, a number of Rett variants have been described. MECP2 gene mutations are responsible for about 90% of classic cases and for a lower percentage of variant cases. Recently, CDKL5 mutations have been identified in the early onset seizures variant and other atypical Rett patients. While the high percentage of MECP2 mutations in classic patients supports the hypothesis of a single disease gene, the low frequency of mutated variant cases suggests genetic heterogeneity. Since 1998, we have performed clinical evaluation and molecular analysis of a large number of Italian Rett patients. The Italian Rett Syndrome (RTT) database has been developed to share data and samples of our RTT collection with the scientific community (www.biobank.unisi.it). This is the first RTT database that has been connected with a biobank. It allows the user to immediately visualize the list of available RTT samples and, using the “Search by” tool, to rapidly select those with specific clinical and molecular features. By contacting bank curators, users can request the samples of interest for their studies. This database encourages collaboration projects with clinicians and researchers from around the world and provides important resources that will help to better define the pathogenic mechanisms underlying Rett syndrome. Hum Mutat 28(4), 329–335, 2007. © 2006 Wiley-Liss, Inc.

KEY WORDS: RTT; Rett syndrome; MECP2; CDKL5; molecular database; clinical database; biobank; biological samples

INTRODUCTION

Rett Syndrome (RTT; MIM# 312750) is a neurodevelopmental disorder that predominantly affects girls and is the second most common cause, after Down syndrome, of severe mental retardation in females [Hagberg, 1995]. It has an incidence of approximately 1 out of 10,000 live female births [Leonard et al., 1997]. RTT patients show a well-defined clinical course and peculiar characteristics. Clinical criteria for the diagnosis of RTT were defined in the 1980’s [Hagberg et al., 1985; Trevathan and Moser, 1988] and recently revised in 2001 to clarify previous ambiguities in interpretation of clinical features [Hagberg, 2002]. In classical RTT girls, birth and early development appear to be normal, although several investigators consider RTT to be a developmental disorder manifesting very soon after birth [Einspieler et al., 2005; Kerr, 1995]. After this apparently normal period, the clinical course is characterized by a stagnation of development followed by regression lasting for several months and usually occurring between 1 to 3 years of age. The fully developed clinical picture is dominated by mental retardation with autistic features, reduction of communication skills, loss of purposeful hand movements combined with hand stereotypes, progressive postnatal microcephaly, abnormal locomotion, and other various signs such as seizures, breathing abnormalities, and other autonomic dysfunctions. In classic RTT, it is possible to appreciate variability in disease severity. Furthermore, several RTT variants have been described, including the “preserved speech variant” (PSV), characterized by preservation of some degree of speech [Fukuda et al., 2005; Yamashita et al., 2001; Zappella, 1992; Zappella et al., 1998]; the “congenital variant,” recognized from birth; the “early onset seizures variant” with seizures onset before regression; the “forme fruste,” with a milder, incomplete clinical course (regression between 1 and 3 years); and the “late regression variant.” Patients showing a striking preservation of their abilities in comparison with PSV girls, especially concerning language and hand use levels, have been reported and classified as highly functioning PSV [Zappella et al., 2003].

Up to 90 to 95% of classic RTT and 40 to 50% of RTT variants are caused by mutations in the X-linked MECP2 gene encoding for

The Supplementary Material referred to in this article can be accessed at http://www.interscience.wiley.com/jpages/1059-7794/suppmat.

Received 16 June 2006; accepted revised manuscript 25 October 2006.

*Correspondence to: Alessandra Renieri, MD, PhD, Medical Genetics, Molecular Biology Department, University of Siena, V.le Bracci, 53100 – Siena, Italy. E-mail: renieri@unisi.it

Grant sponsor: Emma and Ernesto Rulfo Foundation; Grant sponsor: Telethon Foundation; Grant numbers: GGP02006 and GGP05005; Grant sponsor: MIUR; Grant numbers: FIRB 01 and PRIN 2005; Grant sponsor: University of Siena; Grant numbers: PAR 2001, PAR 2002 and PAR 2004.

DOI:10.1002/humu.20453

Published online 21 December 2006 in Wiley InterScience (www.interscience.wiley.com).
methyl CpG binding protein 2 (MIM # 300005) [Weaving et al., 2005]. Recently some groups [Evans et al., 2005; Mari et al., 2005; Nectoux et al., 2006; Scala et al., 2005; Weaving et al., 2005] identified mutations in the CDKL5 gene encoding for cyclin dependent kinase-like 5 (MIM # 300203) in patients with the diagnosis of the early onset seizures variant and in other phenotypes overlapping with RTT. While the high percentage of MECP2-positive classic RTT patients supports the hypothesis of a single gene causing this phenotype, the low frequency of “solved” RTT variant patients supports genetic heterogeneity in RTT variants, as already demonstrated by CDKL5-mutated cases.

To date, three RTT databases have been developed: one established at the University of Edinburgh and the other two funded by the International Rett Syndrome Association (IRSA). Originally developed as a MECP2 mutation collection, the first database has been successively improved with clinical data to allow genotype–phenotype correlations (www.mecp2.org.uk). In 2001, RettBASE was established to collect both published and unpublished data about MECP2 pathogenic mutations, benign polymorphisms, and sequence variations of uncertain significance from around the world (mecp2.chw.edu.au) [Christodoulou et al., 2003]. Two years later, an international RTT clinical phenotype database, linked to RettBASE, became available (www.ichr.uwa.edu.au/rett/irsa) [Fyfe et al., 2003]. All these databases are excellent electronic tools that are very useful for genotype–phenotype correlations. However, none of them is connected to a biobank of RTT patients.

Since 1998, the Medical Genetics Unit of the University Hospital of Siena has collected DNA and lymphoblastoid cell lines from a large number of RTT patients. During the last years, the collection of samples has significantly expanded. The rapid enlargement of the samples collection required the establishment of an online database for data management and sharing of resources with the scientific community (www.biobank.unisi.it). At present (September 2006), the site contains 221 entries corresponding to RTT patients included in the bank. For each entry, the site contains clinical and molecular information about each proband and indicates the biological samples available for patients and relatives included in the bank. The database is a useful tool for researchers working on RTT since they can rapidly search for RTT patients with specific clinical and/or molecular features and, by contacting the bank curators, they can request the biological samples for their studies.

**RTT Bank**

At present (September 2006), the bank contains 631 DNA samples (219 RTT patients and 412 relatives), 49 lymphoblastoid cell lines, and 52 leukocytes in dimethyl sulfoxide (DMSO) medium for a total of 63 probands, and 13 lymphoblastoid cell lines and 61 leukocytes of relatives, also in DMSO medium.

Concerning the phenotype, among the total of 223 probands, 126 are classified as classic RTT and 61 as RTT variants, according to the international diagnostic criteria [Hagberg et al., 2002]. Among RTT variant cases, 41 are PSV, three are “highly functioning PSV,” 10 are early onset seizures variants, five are “forme fruste,” and two are congenital variants. Among the remaining 36 patients, 18 have been classified as RTT-like, i.e., cases who do not completely fulfill the international clinical criteria for RTT, and 18 as not determinable (ND), when the very young age of the patient does not allow a definitive clinical classification.

MECP2 mutations have been identified in 113 out of 126 classic cases (mutation detection rate: 90%), in 27 out of 61 variant cases (mutation detection rate: 44%), in 17 out of 18 ND cases, and in five RTT-like patients. CDKL5 mutations have been found in 4 out of 10 early onset seizures variant RTT patients. In the other 57 cases (13 with a classic phenotype, 30 with variants, one ND, and 13 RTT-like patients), MECP2 and/or CDKL5 mutation screening failed to identify any pathogenic change.

EDTA peripheral blood samples are used for DNA extraction using a QIAamp DNA blood kit (Qiagen, Hilden, Germany; www.qiagen.com). After the extraction, the quality of the DNA is tested through a spectrophotometer. For each sample an aliquot containing at least 400 μg with an Optical Density (OD) 260/280 ratio of 1.8-2 is stored in dedicated boxes at −20°C for long-term conservation. Heparin peripheral blood samples are treated for leukocyte isolation in DMSO medium or establishment of Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines. Four aliquots of transformed cell lines and primary leukocytes are stored at −135°C. Two additional aliquots of cell line of each patient are conserved in a liquid nitrogen dewar to avoid the loss of samples in case of electronic damage. Cells are frozen at passage 2 in 1.5–2 × 10⁷ cells/ml aliquots.

All cases contained in the bank come from all over Italy and all have been clinically evaluated by the Medical Genetics Unit (University of Siena). Patients fulfilling the clinical criteria for RTT and its variants and those who have a mutation in either MECP2 or CDKL5 are inserted in the database. MECP2 and CDKL5 mutation analysis is performed by a combination of DHPLC for all coding exons (exons 1–4) and real-time quantitative PCR (qPCR).

To protect the patient’s privacy, when a biological sample arrives it is stripped of personal identifiers and for each sample an internal code is assigned: RETT followed by a sequential number. Anonymity of the samples is assured in each step of the research. A consent form has to be signed by the patient’s guardians in order to insert the sample in the bank, according to international standards.

**RTT Database**

The RTT database has been available (at www.biobank.unisi.it) since 2004. It is maintained and updated every 3 months on the University of Siena server. The website is written in VBScript and takes advantage of an Internet Information Server (Microsoft IIS; Microsoft, Redmond, WA) with Active Server Pages (ASP) (Microsoft) technology. The website contains an Access (Microsoft) database to manage the data.

In the general homepage of the database, there are links to four independent databases: X-Linked Mental Retardation (XLMR); Rett syndrome; Retinoblastoma; and Other. All these databases are managed by the Medical Genetics Unit (Fig 1). The first two databases are funded through a Telethon grant. By accessing to the RTT section of the bank, users can see the complete list of patients contained in the bank.

The general homepage contains useful links important for site navigation for all the four databases: 1) Guidelines, containing a description of the procedures to follow in order to store and request biological samples (the first option is only possible for the XLMR database); 2) Contact information, for contacting the bank administrator; 3) Services offered from the bank; in particular, isolation of leukocytes from human peripheral blood samples, establishment of EBV-transformed lymphoblastoid cell lines from human peripheral blood leukocytes, DNA extraction, cryopreservation of transformed cell lines and primary leukocytes at −135°C, storage of DNA and plasma at −20°C, and, finally, distribution of the stored biological samples (the latter is the only service available for the XLMR database).
available for the Rett database; 4) Bank organization, describing the general organization of the bank and its sections; 5) Forms, containing the forms users have to complete to take advantage of bank services and the informed consent.

**Database Organization**

The database is organized on two levels: a “public” level freely available to the general public and a “curator” level accessible only to bank curators through the use of a username and password and containing personal data of patients and relatives and their detailed clinical information. All users accessing the main page of the site (Supplementary Fig. S1; available online at http://www.interscience.wiley.com/jpages/1059-7794/suppmat), can visualize a table with the following information: 1) Code: a progressive number that identifies the single families, usually consisting of the affected girl, the parents, and other relatives when available. The system assigns a progressive number automatically every time a new case is inserted, and prevents a double insertion of the same case. 2) Internal code: the personal code assigned to each individual (proband, parents, other relatives) coming to our attention. These codes allow single individuals to be identified without using the personal data. Together with the personal code, the relationship within the family (proband, father, mother, etc.) is visible in this column, so that users can immediately know the individuals with available biological samples. 3) Phenotype category: the phenotypic definition of the proband. Seven phenotypic definitions are available: classic RTT; PSV, which is the most common RTT variant in our series; highly functioning PSV; forme fruste; congenital variant; early onset seizures variant; and Rett-like, including patients with a suggestive phenotype but who can not be framed in one of the specific phenotypes reported above. In addition, for very young patients for whom a definitive clinical classification is not possible the phenotype is indicated as “ND age less than 4 years.” 4) Gene name: the name of the mutated gene in each patient (MECP2 or CDKL5). In those patients for whom the causative mutation has not been identified, the gene name is designed as “Unknown.” 5) Mutation type: pathogenic mutations are classified in four general categories: a) missense mutations, in which a single amino acid has been substituted with a different one; b) early truncating mutations, which interrupt the protein, eliminating part or all of the methyl-CpG-binding domain (MBD) and/or transcription repression domain (TRD) domain; c) late truncating mutations, which interrupt the protein in the C-terminal portion after the TRD domain; and d) gene deletion, either partial or total. 6) Nucleotide change: the change at nucleotide level is reported according to the standard nomenclature [den Dunnen and Antonarakis, 2000]. 7) Amino acid change: the change at protein level according to the standard nomenclature [den Dunnen and Antonarakis, 2000]. 8) Additional info: an icon in this column links to another page containing additional information about the family. On this page, bank curators can visualize all available information about the family, including personal and clinical data of patients and relatives. At the moment, external users can visualize only information about the X-inactivation status of the patient (skewed or partially skewed, with the cut off being 90% and 70%, respectively) [Sharp et al., 2000] and the mutation inheritance (de novo mutation, apparently sporadic, carrier mother, mosaicism in one parent). 9) Biological sample available: a list of the biological samples available for each family (lymphoblastoid cell line, leukocytes in DMSO medium, DNA, fibroblasts). In addition to the type of biological sample, the internal code of all family members for whom that sample is available is reported. Clicking on each sample, only bank curators, through the use of a protected password, can access to the information regarding the location of the sample. 10)
MECP2 and CDKL5 Mutations and Variants

On the main page, in addition to the list of all inserted patients, there are three links in the left upper part of the page, having to do with mutations and variants of MECP2 and CDKL5 genes (Supplementary Fig. S1).

These links are: 1) “List of Mutations”, which provides a table of all identified mutations ordered by their frequency (Supplementary Fig. S2). For each mutation there is the mutation type column, in which the following fields may be present: missense, early truncating, late truncating or gene deletion. The mutations are identified by a systematic name of nucleotide change (GenBank accession number NM_004992 and NM_001037343) and amino acid change (GenBank accession number NP_004983 and NP_003150). One column reports the number of patients in whom the mutation has been found. Another column describes the phenotype of each patient carrying a specific mutation. The last column is dedicated to references, which allows the user to view the PubMed references of papers on the selected patients.

Data in this table are automatically updated every time a new mutated patient is inserted in the database. 2) “Graph of Mutations”, which contains a dynamic graph showing the position of the pathogenic mutations and their relative frequency (Fig 2). On the X-axis of the graph a scaled schematic representation of the protein with the main identified domains is given. On the Y-axis, the frequency expressed as the percentage of mutated patients bearing a specific mutation is given. Mutations are positioned along the protein scheme according to their amino acid numbering and are represented as vertical columns; the height of the column represents the frequency of the mutation. Moreover, from this page, a link allows access to a pie chart that reports the frequency of the different types of mutations (missense, early truncating, late truncating) expressed as a percentage of mutated patients bearing a specific mutation type.

Both graphs are managed by a dedicated program that reads data in the table of mutations and updates the graphs every time a sample is added. At present, MECP2 gross rearrangements are not counted in the graph of mutation and in the pie chart. 3) “Rare variants”, which contains a table listing all nonpathogenic rare variants identified in the patients of the bank (Supplementary Fig. S3). The table includes the following information: a) ID, a progressive number that identifies the single variant; b) nucleotide change; c) amino acid change; d) number of unrelated samples, i.e. the number of unrelated individuals where the variant has been identified; e) samples, i.e. the internal code of all the individuals in which the variant is present; f) reference, which allows the user to view the PubMed references of papers reporting the specific variant.

“Search by” Tool

The website is interactive, with a user-friendly graphical interface. Users that visit the online database can search for a patient by simply looking through the list of registered cases. Alternatively, the database may be explored using the “Search by” tool located in the upper right part of the main page. The search page allows the user to browse the database selecting for specific fields (Fig 3). The patients of interest can be selected by: 1) mutated gene (MECP2 or CDKL5) or absence of mutation (“Unknown” option); 2) mutation type; 3) nucleotide change; 4) amino acid change; 5) phenotype category; and 6) survival, which allows to search alive patients of specific ages or dead patients. Users can choose to search by one single option or to combine two or more options; e.g. they can search all patients with missense mutations or they can search all patients with missense mutations in the MECP2 gene and a PSV phenotype. As a result of the search, they will visualize a table containing all the fields present in the general table and listing all patients fulfilling the requested characteristics. This extremely flexible search option allows users to perform an accurate selection of patients so that they can immediately evaluate whether the type of samples they need are present in the bank and can choose the samples to request before contacting the bank curators.

Security and Quality Assurance

The database has been constructed in compliance with the guidelines of the Italian Society of Human Genetics and Telethon Foundation for biobanking [Dagna Bricarelli et al., 2003]. The database design assures patients’ anonymity, privacy, and confidentiality, according to international criteria [Godard et al., 2003]. Moreover, the informed consent has been written explicitly, covering all aspects of stored samples and personal data management [Godard et al., 2003].

The RTT bank is available to all users who contact bank curators and fill in and sign a specific form asking for biological samples. Biological samples stored in the bank are distributed only to qualified professionals for research purposes only. The specimen cannot be distributed to other investigators without previous written permission of the bank curator. Researchers are asked to include the name of one of our researchers and/or to acknowledge the bank in any paper that includes results obtained using the bank samples/services. Anonymity of the samples is assured in each step of the research. Only the bank curator has access to the whole information content of the database, including personal data.

DISCUSSION

The Italian RTT database has been developed to share data and samples of our RTT biobank with the scientific community. The database allows to immediately visualize RTT samples contained in the biobank and to rapidly select those with specific clinical and molecular features. Contacting bank curators, users can request the samples of interest, following simple procedures indicated on the homepage of the website. In comparison with the already existing RTT databases, this database is the first one that is connected with a biobank. It is important to underline that all patients inserted in the database are evaluated by the same group of clinicians and the molecular analysis is performed by the same laboratory, allowing a uniform clinical and molecular data collection. The database includes both classic and variant RTT patients and cases who do not completely fulfill the international clinical criteria for RTT, but who have clinical features strongly resembling the RTT phenotype (RTT-like patients). Our cohort of MECP2-mutated patients also comprises those patients with a MECP2 gene deletion. Using the “Search by” tool, users can select
for patients with a MECP2 mutation or a CDKL5 mutation, or for patients in whom the molecular analysis failed to reveal a pathogenic mutation. The well-characterized collection of mutated patients is very important for researchers working on genetic modifiers of the RTT phenotype or on the functional consequences of a specific mutation, while mutation-negative RTT patients are potentially important for the identification of other genes involved in the syndrome.

During the last years, it has become evident the importance of international data to perform accurate genotype–phenotype studies [Leonard et al., 2005; Robertson et al., 2006]. InterRett, linked to RettBASE, has collected data on a large scale sample of

![Graphical view of mutation distribution within the MECP2 (a) and CDKL5 (b) gene.](www.interscience.wiley.com)
cases (2,089) representing 28 countries around the world and it is the most powerful tool for researchers working on genotype-phenotype correlations. For this reason, since our database contains a collection of RTT patients limited to the Italian population, we will make all effort to support this international project. However, it is important to also maintain and to improve this database to preserve its unique features. In addition to the already-mentioned characteristics, our database contains data on the X-inactivation status for a great portion of patients; these data are freely available for the general public. This has been possible thanks to the collection of parents’ DNA samples. Since one of the most important factors in phenotype modulation is the X-inactivation status, this information will allow users to perform more accurate genotype-phenotype correlations once the patients’ clinical section is improved (see Future Prospects). In addition, this information is fundamental for researchers who request biobank samples for expression studies.

In a separated section, the database contains tables indicating MECP2 and CDKL5 polymorphisms/rare variants with the code for the patients in whom they have been identified. This information, together with detailed clinical data (see Future Prospects), is important to study if these variants contribute in modulating the phenotype. MECP2 and CDKL5 interact suggests that the association between mutations in one of the two genes and specific variants in the other can modulate the RTT phenotype [Mari et al., 2005].

Given that for rare disorders such as RTT, the Internet provides an important means of communication, this online database encourages collaborative projects with clinicians and researchers from all around the world. This database, providing the opportunity for researchers to take advantage of this collection of clinically and molecularly well-characterized patients, represents an important resource to accelerate the clarification of the molecular basis of RTT.

**FUTURE PROSPECTS**

A detailed clinical characterization of each patient is a fundamental step for inclusion in the database. At present, all clinical details except for the general classification in classic, variant, or RTT-like cases, are available only for bank curators on the “Additional info”, page. We are planning to create, for each case, a table with a set of clinical features and a related clinical score according to data in the literature [Charman et al., 2005; Colvin et al., 2003; Huppke et al., 2002; Kerr et al., 2001; Monros et al., 2001]. The new schedule will be freely available to the general public. Users will access this information by clicking on “Additional info”, which already exists on the main page. This amelioration will make possible a better definition of the clinical phenotype of each patient. All these clinical features will be added in the “Search by” tool, making it possible for all users to sort cases by particular fields in the clinical data.

Regarding the X-inactivation data, we plan to ascertain the X-inactivation status of all patients included in the database and to display this information on the “Additional info” page.

Furthermore, the bank includes RTT patients with chromosomal rearrangements not involving MECP2 or CDKL5 [Delobel et al., 1998; Pescucci et al., 2003]. The expanding number of these cases will eventually lead to the identification of patients with overlapping rearrangements. These cases will possibly contribute to the identification of new RTT candidate genes. We plan to add the column “chromosomal rearrangements” to the main page of the database in order to allow users to immediately visualize this information and, eventually, to request the samples.
ACKNOWLEDGMENTS

We thank the members of the RTT families for their important role in this work. We also acknowledge the Telethon Foundation for the GGP02006 and GGP05005 grants to A.R. This work was also supported by the University of Siena (PAR 2001, PAR 2002, and PAR 2004), and by MIUR (PRIN 2006) to A.R.

REFERENCES


