Published in final edited form as:

Genesis. 2009 December ; 47(12): 842–846. doi:10.1002/dvg.20575.

Textpresso Site-Specific Recombinases: a text-mining server for the recombinase literature including Cre mice and conditional alleles

William M. Urbanski^{1,2} and Brian G. Condie^{1,*}

¹Department of Genetics, Developmental Biology Group, University of Georgia, Athens, GA 30602

²Department of Computer Science, Developmental Biology Group, University of Georgia, Athens, GA 30602

Abstract

Textpresso Site Specific Recombinases (<http://ssrc.genetics.uga.edu/>) is a text-mining web server for searching a database of over 9000 full-text publications. The papers and abstracts in this database represent a wide range of topics related to site-specific recombinase (SSR) research tools. Included in the database are most of the papers that report the characterization or use of mouse strains that express Cre recombinase as well as papers that describe or analyze mouse lines that carry conditional (floxed) alleles or SSR-activated transgenes/knockins. The database also includes reports describing SSR-based cloning methods such as the Gateway or the Creator systems, papers reporting the development or use of SSR-based tools in systems such as *Drosophila*, bacteria, parasites, stem cells, yeast, plants, zebrafish and *Xenopus* as well as publications that describe the biochemistry, genetics or molecular structure of the SSRs themselves. Textpresso Site Specific Recombinases is the only comprehensive text-mining resource available for the literature describing the biology and technical applications of site-specific recombinases.

Keywords

Site-specific recombinase; Cre; FLP; PhiC31; Bxb1; Cre mouse database

Introduction

Site-specific recombinases (SSRs) such as Cre and FLP have become invaluable genetic tools ((Branda and Dymecki, 2004; Calos, 2006; Furuta and Behringer, 2005; García-Otín and Guillou, 2006; Gilbertson, 2003; Hadjantonakis *et al.*, 2008; Khan *et al.*, 2005; Kim and Dymecki, 2009; Lee and Luo, 2001; Lewandoski, 2007; Marsischky and Lobaer, 2004; McGuire *et al.*, 2004; Mills and Bradley, 2001; Nagy *et al.*, 2003; Ow, 2002, 2007; Venken and Bellen, 2005, 2007). SSRs have been adapted for use in a wide variety of different

* Author for Correspondence: Brian G. Condie Ph.D., 270A Coverdell Center, Department of Genetics, University of Georgia, Phone: 706-543-1431, Fax: 706-583-0691, bcondie@uga.edu.

organisms and experimental systems and SSR-based techniques have been developed for a wide range of applications. Also, the genetics, biochemistry and molecular structures of several widely used SSRs have been reported. As a result, the scientific literature describing these recombinases and the research methods that use them has become very large, complex and diverse. It is now a challenge for investigators to identify site-specific recombinase research tools that are appropriate for their research.

Because of the diversity and size of this literature, we have developed a Textpresso Site-Specific Recombinases (Textpresso SSR) web server to provide a text mining tool for researchers to more easily retrieve published information about SSRs and SSR-based technologies. Textpresso SSR searches a database that contains most of the papers describing mouse strains that express Cre recombinase and strains that carry conditional alleles or transgenes that are activated or inactivated by Cre activity (Table 1). Textpresso SSR can be used to identify these strains and it is a unique tool for recovering specific pieces of information from the published results of tissue specific gene knockouts in mice. Also, the database contains papers from many other areas (Table 1). Recombinases included in the database are Cre, Flp, PhiC31, Bxb1 and λ int.

Results and Discussion

The Textpresso SSR database contains a wide range of publications

As of September 2009 the Textpresso SSR database contained 9326 publications. Table 1 illustrates the major topics in the database. Please note that the topics listed in Table 1 are the major areas in the database and do not represent all of the content of the database.

We anticipated that a major use of this web server would be to retrieve information about mouse strains including strains that express Cre recombinase, mice that carry conditional alleles (floxed alleles) and mice with transgenes that can be activated or inactivated by SSRs. We made every effort to include most of the existing publications that describe these mouse resources. Because we have included nearly all of the full text publications or PubMed abstracts that describe SSR-based mouse resources, Textpresso SSR can be used to retrieve information about the phenotypic consequences of tissue-specific or temporally controlled gene knockouts. Thus, Textpresso SSR is a unique resource for searching this very important segment of the mouse genetics literature.

We also included publications that describe the properties of SSRs as well as applications in other organisms. For example, SSRs are now widely used in *Drosophila* and in several popular molecular cloning systems. Investigators who work with any of the systems listed in Table 1 will find Textpresso SSR to be a rich information source for SSR based research tools as well as the results obtained with them. In addition, the publications describing the biochemistry, genetics or structures of Cre, Flp, PhiC31 and Bxb1 recombinases are included in the database.

Input and Output

The Textpresso SSR web server presents the user with a standard Textpresso search page (Figure 1A). Navigation between pages is facilitated via the buttons (links) at the top of each

page. We have provided a FAQ and User Guide page to help address common user questions as well to explain key features of the search interface. The FAQ and User Guide provides a great deal of specific information and guidance on effective search strategies for users who want to perform extensive searches. We strongly recommend that new users carefully read the FAQ and User Guide page on the web site before performing any searches.

The user has several search options. The database can be searched with keywords alone, keywords plus one to four categories or one to four categories without keywords. Multiple keywords can be used in a single search. Spaces between keywords are recognized as the Boolean operator “AND” while commas with no spaces are recognized as “OR”. The categories are selected from pull down menus on the search page (Figure 1A). Each category is linked to an ontology that contains the terms and concepts that fall within that category. Users can enter a keyword into the ontology browser to identify the category that contains the term. The ontology browser is available via the “Ontology” button at the top of each Textpresso SSR page (Figure 1A). The search page also lets the user set options for the fields within the database that are to be searched as well as the search scope, search mode and the way the search results are sorted in the output (Figure 1A). These functions have been described in detail previously (Müller *et al.*, 2004).

It is likely that many Textpresso SSR users will be seeking information about mice that express Cre recombinase or for information about conditional alleles of mouse genes. Publications that present this data can be easily found within the database by searching with specific keywords. In the search shown in Figure 1A, the keywords “foxn1 cre” were entered into the keyword box and exact match was selected from the Keyword Specification options. Since these keywords were separated by a space the search was actually performed by Textpresso using the query “foxn1 AND cre”. This keyword query returned 13 matching papers, one of these matches is shown (Figure 1B). In cases where a gene is known by multiple names, the user can create a keyword query that will retrieve publications containing the various synonymous gene names. Gene name synonyms can be obtained from the MGI web server or from the Gene Ontology website. For Cre transgenics or knockin mice the MGI Phenotypes, Alleles and Disease Models Query page can be used to find synonymous names for a particular transgenic or knockin mouse line. Detailed instructions for performing this type of search can be found on the FAQ and User Guide page of Textpresso SSR.

The other commonly used search method is to perform a search with keyword(s) and categories. Categories are collections of related terms and concepts grouped under a category heading such as “developmental process” or “disease (H. sapiens)”. This kind of search has the advantage that it will return a large number of terms that are within the category and that appear in sentences with the keyword or keywords. This can be useful in cases where the user wants to do a very broad search and several terms within the category are of interest. Using a category search is also helpful when the user is searching for publications in a category that is rich in synonymous terms.

Validation and Testing of Textpresso SSR

We have performed several tests of Textpresso SSR. These have been designed to test the ability of the Textpresso search engine to retrieve papers from a set of publications that are known to be within the text database. In the two cases where we tested the author search function (with exact match selected under keyword specification) we were able to retrieve 100% of the publications that matched the each of the authors with no irrelevant publications retrieved. This suggests that author searches within the database will have very high retrieval and precision rates.

We performed a second series of tests to test the ability of Textpresso retrieve publications that describe or use specific lines of Cre expressing mice. We chose to test the ability of Textpresso SSR to retrieve these papers because we have a record of the publications that were uploaded into the database. This record is based on the publications listed by the MGI database as corresponding to each strain at the time of the most recent update of the publication database of Textpresso SSR. The record from MGI gives us a reference set to compare our search results to, allowing us to calculate a retrieval efficiency.

We searched for three different Cre strains: Synapsin1-Cre, Meox2-cre and Twist2-cre. In the case of Synapsin1, we used two separate searches (see the Methods for details). Together the searches retrieved 28 out of the 30 of the papers known to be in the database, giving a retrieval rate of 93%. For Twist2-cre we also used two different search queries that retrieved 20 of the 21 papers known to be in the database (95% retrieval). With Meox2-Cre we also used two separate searches and Textpresso retrieved 75 out of 92 references (82%). Overall, these studies suggest that Textpresso is very efficient at retrieving publications that describe Cre expressing mice, when appropriate keyword queries are used.

In each test search, Textpresso returned additional references beyond the ones listed in MGI record. We examined these additional, potentially irrelevant publications in detail for the Twist2-Cre searches. We found that out of 19 references that were not listed on the MGI record, 12 did contain a reference to the Twist2-Cre strain in the text of the paper. Each of these contained information relevant to the Twist2-Cre strain and are therefore relevant to the query. We then examined the 7 remaining potentially irrelevant search results. Two of these papers each reported studies that used the Twist2-Cre mice but had been missed by the MGI curators. One other publication was a review that had not been listed on the MGI record for Twist2 and another paper reported the characterization of a Twist1-Cre mouse. The 3 remaining papers mentioned the Twist genes in terms of their expression in the context of the experiments reported in each paper. This analysis suggests that Textpresso SSR is capable of returning nearly all of the publications that describe a particular Cre expressing mouse strain as well as additional information that is relevant to the user.

Methods

Database construction

Our goal was to build a text database for a wide range of researchers. A focus was to acquire most of the publications that describe mouse strains that express SSRs as well as most of the papers that describe mice with conditional alleles that can be inactivated or activated in

response to SSRs. To do this, we used the MGI database at the Jackson Laboratory as a source of pmid numbers for publications that contain *any* information about relevant mice. All of the available publications that correspond to these strains were downloaded. In addition, we used broad PubMed searches using the keywords “recombinase”, “Cre”, “Flp”, “PhiC31”, “Bxb1”, “integrase” and “lambda integrase” were performed to acquire a broad set of references that describe these SSRs. Additional searches were performed using the MESH terms “Cre recombinase “[Substance Name] and “FLP recombinase “[Substance Name]. We also included a number of publications that describe the Gateway and Creator cloning systems as well as many of the genomic and proteomic resources that have been generated using these systems. These publications were downloaded from PubMed searches using “Creator” or “Gateway” as keywords and were obtained from the Gateway publication lists provided on the Invitrogen website. Also, we downloaded papers referenced in comprehensive review articles describing SSR based research resources.

To build the publication database we downloaded pdf files using the program Papers (<http://mekentosj.com/>). Although it is possible to import pdf files directly into a Textpresso database, we decided to use Papers so that we could assemble the database on laptop computers and perform manual curation. This step screened out publications and abstracts that were retrieved because of an irrelevant match between the publication and the search term and it removed papers that were outside of the intended scope of the database. The pdf files and the PubMed citation information and abstracts for the publications were imported into Textpresso. In cases where a pdf file was not available to us, we inserted the PubMed abstract into the Textpresso SSR database. After construction of the initial build of the publication database, the Textpresso SSR server was tested on four different web browsers and several different operating systems: Firefox (Mac, Windows and Linux), Safari (Mac), Internet Explorer (Windows) and SeaMonkey (Mac). We will update Textpresso SSR at approximately 6 week intervals. These updates will be performed manually to ensure that only relevant publications are added to the database.

The Textpresso software has been described in detail previously (Müller *et al.*, 2004). To build the Textpresso SSR web server and publication database we used the open source version of Textpresso 2.0 and the Alere 0.1.0 scripts that are available for download from the central Textpresso website (<http://www.textpresso.org/downloads.html>).

Testing Textpresso SSR

We performed searches for references that describe or mention the Twist2-Cre, Meox2-Cre or Synapsin1-Cre mouse strains. In the case of Twist2-Cre one search was performed using the keyword string Dermocre,Dermo1-Cre,Dermo1cre,twist-2-,Twist2tm1(cre)Eno and the second search was performed with the keyword twist2. The synonymous strain names for the various Twist2 mice used to generate the keyword string were obtained from an MGI Phenotypes Alleles and Disease Models search. Detailed instructions for constructing this type of query are provided on the FAQ and User Guide page. We also performed two searches for Meox2-Cre mice. The first search used the keyword string meox,Meox2-Cre,Meox2Cre,Meox2CreSor, More-cre,Mox2 Cre,Mox2-Cre,Mox2Cre and the second search used the keyword meox2. For the Synapsin1-Cre strain we also performed two

searches, one using the keyword string *synapsin1-cre,synapsinI-cre,Syn-cre,SynI-Cre,synapsin-cre,syn1-cre* while the second search used the keyword *synapsin cre*.

Acknowledgements

The authors thank Nathan Bond and Patricia Condie-Brown for their assistance in building and updating the literature corpus that is the basis of Textpresso SSR. We thank Ruihua Fang and Hans-Michael Muller at Cal Tech for their advice and assistance with the Textpresso software.

Funding

This work was supported by the University of Georgia Office of Vice President for Research.

Literature Cited

- Branda CS, Dymecki SM. Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice. *Dev Cell*. 2004; 6:7–28. [PubMed: 14723844]
- Calos MP. The ϕ C31 integrase system for gene therapy. *Current gene therapy*. 2006; 6:633–645. [PubMed: 17168696]
- Furuta Y, Behringer RR. Recent innovations in tissue-specific gene modifications in the mouse. *Birth Defects Res C Embryo Today*. 2005; 75:43–57. [PubMed: 15838923]
- García-Otín AL, Guillou F. Mammalian genome targeting using site-specific recombinases. *Front Biosci*. 2006; 11:1108–1136. [PubMed: 16146801]
- Gilbertson L. Cre-lox recombination: Cre-ative tools for plant biotechnology. *Trends Biotechnol*. 2003; 21:550–555. [PubMed: 14624864]
- Hadjantonakis AK, Pirity M, Nagy. Cre recombinase mediated alterations of the mouse genome using embryonic stem cells. *Methods Mol Biol*. 2008; 461:111–132. [PubMed: 19030793]
- Khan MS, Khalid AM, Malik KA. Phage ϕ C31 integrase: a new tool in plastid genome engineering. *Trends Plant Sci*. 2005; 10:1–3. [PubMed: 15642516]
- Kim JC, Dymecki SM. Genetic fate-mapping approaches: new means to explore the embryonic origins of the cochlear nucleus. *Methods Mol Biol*. 2009; 493:65–85. [PubMed: 18839342]
- Lee T, Luo L. Mosaic analysis with a repressible cell marker (MARCM) for *Drosophila* neural development. *Trends Neurosci*. 2001; 24:251–254. [PubMed: 11311363]
- Lewandoski M. Analysis of mouse development with conditional mutagenesis. *Handb Exp Pharmacol*. 2007:235–262. [PubMed: 17203658]
- Marsischky G, Labaer J. Many paths to many clones: a comparative look at high-throughput cloning methods. *Genome Res*. 2004; 14:2020–2028. [PubMed: 15489321]
- McGuire SE, Roman G, Davis RL. Gene expression systems in *Drosophila*: a synthesis of time and space. *Trends Genet*. 2004; 20:384–391. [PubMed: 15262411]
- Mills AA, Bradley A. From mouse to man: generating megabase chromosome rearrangements. *Trends Genet*. 2001; 17:331–339. [PubMed: 11377795]
- Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004; 2:e309. [PubMed: 15383839]
- Nagy A, Perrimon N, Sandmeyer S, Plasterk R. Tailoring the genome: the power of genetic approaches. *Nat Genet*. 2003; 33(Suppl):276–284. [PubMed: 12610537]
- Ow DW. Recombinase-directed plant transformation for the post-genomic era. *Plant Mol Biol*. 2002; 48:183–200. [PubMed: 11860209]
- Ow DW. GM maize from site-specific recombination technology, what next? *Curr Opin Biotechnol*. 2007; 18:115–120. [PubMed: 17353124]
- Venken KJ, Bellen HJ. Emerging technologies for gene manipulation in *Drosophila melanogaster*. *Nat Rev Genet*. 2005; 6:167–178. [PubMed: 15738961]
- Venken KJ, Bellen HJ. Transgenesis upgrades for *Drosophila melanogaster*. *Development*. 2007; 134:3571–3584. [PubMed: 17905790]

Figure 1. URBANSKI AND CONDIE

An example of a keyword search for publications describing mouse strains that express Cre recombinase from the Foxn1 gene or promoter

A. The Textpresso SSR Search Page. Major features are indicated. At the top of the page are buttons that link to all of the pages of the site (blue arrow). The black arrow indicates the pull-down menus for selecting search categories. The green arrow indicates the boxes that allow the user to select which fields within the text database are to be searched. The two display option controls (gray arrow) control the number of sentences displayed that are adjacent to each match to the keyword query (matching sentences) or the number of retrieved publications displayed per page (entries/page). Additional details about the Search Page can be found on the FAQ and User Guide page on the Textpresso SSR website.

B. An example of one of the 13 publications found in this search. The keywords used in the search are highlighted in the relevant sentences from the paper. This search result has been truncated at 8 matching sentences. The presentation of sentences that match a query allows the user to scan the retrieved publications for information of interest.

Table 1
Topic areas represented in the Textpresso SSR publication and abstract database

General topic areas	Specific topics within general topic area
SSR expressing mice.	Characterization of Cre, Flp or PhiC31 expressing mouse strains. Tamoxifen-inducible recombinases. Phenotype analysis of tissue or temporal specific knockouts.
Conditional knockout mice.	Characterization of conditional alleles. Phenotype analysis of tissue or temporal specific knockouts.
Conditional gene activation in mice.	Conditionally activated transgenes or knockin alleles. Cell lineage and cell fate analysis. Rosa26 knockin transgenes, brainbow and related methods.
Emerging site specific recombinase tools.	Feasibility studies of Dre, PhiC31, Bxb1, U153, TP901-1, A118, PhiFC1, PhiRV1 in eukaryotic cells and/or mice.
PhiC31 applications in mammals.	Single copy integration at endogenous att sites. Deletions in mice.
Applications in Drosophila	Production of deletion mutations. Production of chromosomal rearrangements. Mitotic recombination and MARCM. Site-specific transgene integration. Cell lineage marking. Targeted gene knockout.
Yeast applications.	Selection marker removal in multiple gene disruption strategies. Specialized Gateway vectors.
Bacterial applications.	Insertions targeted to att sites. Regulation of exogenous protein over-expression. Selection marker removal.
SSR based molecular cloning and genetic engineering.	Gateway cloning and projects using Gateway cloning tools. Creator cloning. Univector system. Recombineering.
Parasite studies.	Transgenesis in Plasmodium, Conditional mutagenesis in Plasmodium Parasite-host interactions. Database includes studies of Plasmodium, Toxoplasma, Leishmania, Trichinella, Nippostrongylus, Schistosoma, Trypanosoma and Trichuris.
Biochemical, genetic and structural studies of SSRs.	Structural, genetic and biochemical studies of Cre, Flp, Bxb1, Phic31 and λ int. Recombinase protein engineering.
Applications in plant systems.	Removal of selectable markers. Transgene activation. Inducible recombinases. Database includes applications in Arabidopsis, rice, maize, tobacco, orange, wheat, citrus, strawberry, potato and tomato.
Applications in Zebrafish and Xenopus	Cell lineage tracing. Production of transgenics and regulation of transgene expression.
Applications in viral vector production and use.	Adeno-associated virus vectors, Adenovirus vectors and Adeno-Cre virus applications.
Applications in stem cells	RMCE , lineage analysis, gene regulation, cell permeable recombinases, viral gene transfer, PhiC31 mediated integration at endogenous att sites.