RetroScope : A Visualization System for Retro Elements.

정우근
Chung Woo-Keun
부산대학교 컴퓨터공학과
wkchung@pusan.ac.kr

ABSTRACT

최근 전유전체의 비교를 통하여 인간 특이적인 유전자를 밝혀 내고 있는 실정이다. 또한 인접한 유전자간의 관계를 종간 전유전체적으로 분석하고 있으며, Browser[11],[10],[8],[7],[2] 들을 통하여 biologist 또는 사용자에게 제공되고 있다. 우리는 이러한 브라우저, 또는 연구를 통하여 유전자간의 관계를 전유전체적으로 분석하였으나, 이러한 원인이 되는 DNA 재배열과 RetroElement 의 삽입에 대해서는 알 수가 없다. 본 논문에서는 유전체의 45% 를 차지하고 있는 RetroElement 에 대해서 분석하고자 한다. 본 논문에서는 전 Genome 상에 존재하는 Retro Element 데이터 즉, ERV 데이터를 시각화하는 Browser 를 제안한다. 본 논문에서 제안하는 Browser 는 어떠한 환경에도 작동하는 Component Ware 형태로 시스템을 제안한다. 본 논문에서 제안하는 시스템을 RetroScope 부르기로 한다. 본 논문에서 제안한 Component Ware 인 RetroScope 는 인간과 다른 종들간의 유전체내에 많은 유전자와 더불어 많은 Copy 를 차지하는 RetroElement 의 인자만을 특정화 시켜서 보여주며, 각 염색체에 존재하는 RetroElement 에 대하여 비교분석도 가능하도록 제공 하고 있다. 또한 각 개체에서 어떤 패턴으로 Retro Element 가 존재하는지 또한 하나의 염색체 존재하는 Retro Element 를 다중적으로도 볼수있도록 제공하고 있다.

KEYWORDS Genome Browser, RetroElements, ERV, Endogenous Retro Viruses

1 Introduction

전유전체의 비교를 통하여 인간 특이적인 유전자를 밝혀 내고 있는 실정이다. 이러한 유전체들을 비교하고, 특이적인 유전자의 특징을 밝혀내는 일에 큰 몫을 하는 것이 바로 Genome Browser 이다. 이러한 Genome Browser 는 개인 적인 연구, 또는 한 단체에서의 연구에서 쓰이나 공개된 Browser 도 많은 실정이다. 우리는 이렇게 공개된 Browser 를 통하여 인접한 유전자간의 관계를 종간 전유전체적으로 분석하고 있으며 이러한 Browser 를 통해 사용자 또는 biologist 에게 정보를 제공하고 있다. 이렇게 공개된 Browser 를 통하여 우리는 생물정보학에 많은 기여를 하고 있다. 하지만 우리는 이러한 Browser, 또는 연구를 통하여 제공되는 유전자간의 관계를 전유전체적으로 알수 있으나, 이러한 원인이 되는 DNA 재배열과 retroelement 의 삽입에 대해서는 알 수가 없다. 그리하여 본 논문에서는 이러한 원인이 되는 것중 하나인 RetroElement 에 대하여 시각화를 제공하며 분석하고자 한다. 다양한 유전자들이 진화상에서 DNA 재배열에 의해서 gene family 를 형성하거나, 사라지기도 한다. 이러한 재배열의 가장 큰 요인은 염기서열을 유사성을 가진부분에 의한 것이다. 따라서 유전체 전반에 산재해 있고, 염기서열의 유사성이 높은 레트로 엘리먼트와의 위치규명이 필요하다. 실제로 이러한 현상이 일어나곳에서의 경계선은 레트로엘리먼트의 비중이 높다고 알려져있다. 이렇듯 우리는 레트로 엘리먼트와의 위치 규명이 필요한 실정에 임하여 각 염색체에 존재하는 RetroElement 들을 시각화 하고자한다. 본 논문에서 제공되는 RetroScope 는 각 종족, 염색체 마다 존재하는 RetroElement Data 들을 보여주며, 사용자 및 biologist 들의 쉬운 비교 분석을 위하여 Comparative 시각화도 제공하며 있으며, 각 염색체 존재하는 Retro Element Data 들에 한하여 Exon, Intron 과의 Overlapped 되는 부분도 시각화도 제공하고 있다. 이러한 사용자 인터페이스에 관해서는 단락을 통하여 자세히 알아보도록 하겠으며, 다음단락에서는 RetroScope 의 전체적인 System Structure 에 대해서 알아보도록 하겠다.

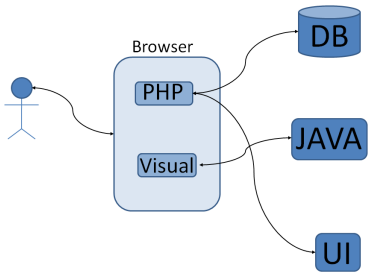


그림 1. 본 논문에서 제시한 RetroScope의 전체적인 구조. 사용자는 PHP 기반으로 구성된 Brower에서 UI를 통하여 시각화를 제공받는다.

2 System Structure

본 단락에서는 RetroScope의 전체적인 시스템 구조, 그리고 데이터의 흐름을 소개한다. 본 단락에서 소개할 RetroScope는 유전체 데이터를 바탕으로 한 Genome Browser이다. RetrpScope에서 제공되는 시각화는 유전체 데이터를 바탕으로 시각화를 제공하는데, 본 논문의 시스템에서는 RetroElement 데이터를 DataBase로 구축하여, RetroScope에서 데이터가 저장된 서버 컴퓨터와의 엑서스를 통한 데이터를 바탕으로 사용자에게 시각화를 구성한다. 본 논문에서 소개된 RetroScope의 효과적인 서비스를 위하여 Component Ware를 기반으로 구성되어 있다. 어떠한 환경에서도 애플리케이션을 구축하여 서비스를 제공할 수 있는 Component ware 환경은 JAVA 언어로 이루어져 있다. JAVA 언어의 플랫폼 독립적 특성때문에 다양한 시스템 환경을 위한 애플리케이션 제작에 JAVA는 좋은 언어가 된다. 플랫폼 기반의 환경은 Component Ware 환경 기반의 애플리케이션 제작에 있어서 웹상에서의 서비스를 위하여 JAVA 언어의 Component 중 하나인 Applet으로 시각화를 제공하기로 하였다. 본 논문에서 제시한 RetroScope를 사용하는 사용자 및 biologist 들의 편의성을 고려하여 본 시스템은 편리한 UI를 제공한다. 브라우저에서 제공되는 UI는 PHP 언어로 구성되어 있다. PHP는 뛰어난 성능을 가지고 있으며, 다양한 DataBase 지원하는 인터페이스, 일반적인 웹 기능을 지원하는 다양한 내장 라이브러리, 호환성이 강한 언어이다. PHP의 이러한 특성을 이용하여 RetroScope는 PHP에서 제공되는 Database 인터페이스를 바탕으로 RetroElement 데이터로 구성된 Database와의 엑서스를 바탕으로 데이터를 추출한다. 그림 **??**, 본 논문에서 제시한 RetroScope의 전체적인 데이트 흐름을 알아보도록 하자. RetroScope는 사용자에게 편리한 UI를 제공한다. 사용자는 편리한 UI를 통하여 보고싶은 데이터를 선택하거나 또한 전체적인 시각화를 제공받을 수 있다. 사용자에게 전체적인 또는 선택적인 데이터를 바탕으로 한 시각화의 제공은 JAVA 언어로 구성한다. 사용자의 선택적인 또는 부분적인 데이터 추출은 PHP에서 제공되는 UI를 통하여 사용자의 명령에 따른 정보에 따라서 질의문을 형성한다. 형성된 질의문을 통하여 유전체 데이터를 바탕으로 이루어진 Database의 질의를 통하여 나온 결과를 바탕으로 JAVA 언어를 통하여 시각화를 제공한다. PHP에서 지원하는 Database 인터페이스를 사용하면 쉽고 간편한 질의문을 형성과 Database 질의문을 전송하여 쉽게 결과를 얻을 수가 있다.

본 단락에서는 RetroScope의 전반적인 시스템 구조와 제공되는 UI 그리고 시각화를 구성하는 언어에 대하여 알아보았다. 다음 단락에서는 RetroScope에서 제공되는 시각화에 대해서 알아보도록 하자.

3 User Interface

본 단락에서는 RetroScope에서 제공되는 UI와 제공되는 시각화가 어떤 것이 있는지에 대해서 알아보도록 하겠다. 본 논문에서 제시한 RetroScope는 기본적으로 각 종족에 존재하는 RetroElement에 대하여 각 염색체에 대하여 시각화를 Comparative하게 제공 하고 있다. 브라우저의 창을 반으로 나뉘어서 같은 정보를 두곳에서

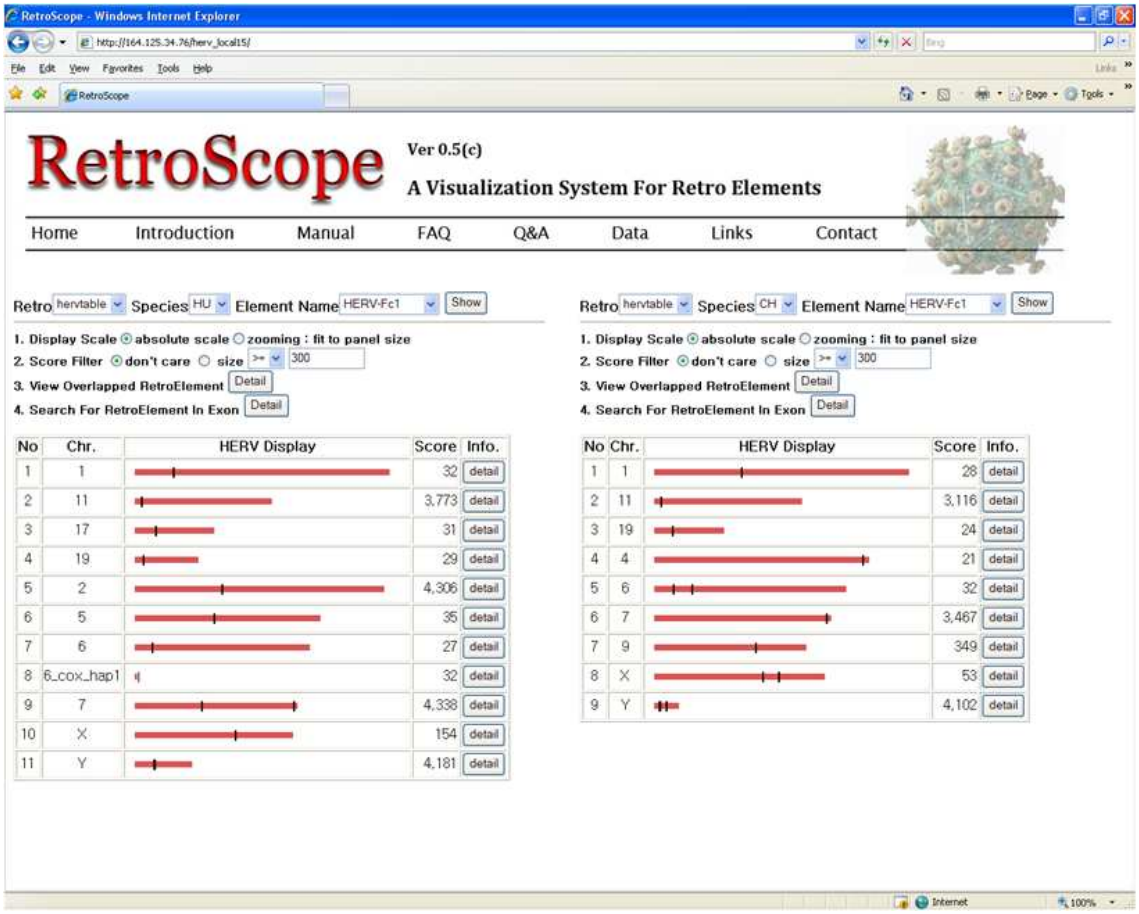


그림 2. 본 논문에서 제시한 RetroScope의 전체적인 모습이다. 데이터 간의 비교를 위하여 Comparative한 시각화를 제공하고 있는 모습이다.

제공하여 사용자로 하여금 좀더 다양한 시각화를 제공하는 것이다. 이러한 기본적인 시각화를 바탕으로 다양한 시각화를 제공 하고 있다. 각 염색체에 존재하는 RetroElement에 대해서 3개 미만의 RetroElement를 선택하여서 해당 데이터들을 중첩시켜서 보여주며, 또한 Exon & Intron 영역에 존재하는 RetroElement들을 조사, 그리고 영역안에 존재하는 RetroElement를 보여준다. 다시 한번 정리하자면 RetroScope에서 제공되는 시각화는 다음과 같다.

- 각 종족, 염색체에 존재하는 RetroElement에 대해서 Comparative한 시각화 제공
- 각 종족, 염색체에 존재하는 Exon & Intron 영역에 존재하는 RetroElement 시각화 제공
- 각 종족, 염색체에 존재하는 RetroElement에 대하여 중첩된 시각화 제공.

위와 같은 시각화에 대해서 자세히 알아보도록 하자. RetroScope에서 제공되는 기본적인 Comparative한 시각화는 2이 되겠다. 상반부에는 홈페이지에서 제공되는 메뉴가 보이며, 그림 3 와 같은, RetroScope에서 제공되는 메뉴도 볼 수 있다. 아래에 존재하는 테이블안의 빨간색 선이 염색체의 길이를 나타내고 있으며, 검은색 선이 RetroElement가 존재하는 영역을 표현하고 있다. 테이블 안에 존재하는 detail 버튼이 그림 6 에 존재하는 좌측 그림와 같은 창을 제공하여, 각 종족, 염색체에 존재하는 Exon & Intron 영역에 존재하는 RetroElement들이 존재하는 지에 대한 여부를 판단하여 데이터가 존재할 경우 버튼을 활성화 하여, 그림 6 와 같은 서비스를 제공한다. 그림 6에 나타는 A는 기본적은 Exon & Intron 영역안에 존재하는 RetroElement들을 시각화 하고 있으

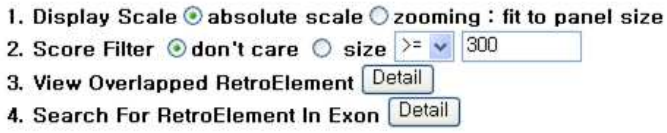


그림 3. RetroScope의 메뉴이다. 1번은 데이터 시각화를 할 경우 절대적인 크기 또는 패널 사이즈에 크기를 맞출수 있는 Option 박스를 보여준다. 2번은 Score 사이즈에 따라서 데이터를 추스려 낼수 있는 기능이다. 3번은 각 염색체 존재하는 RetroElement의 데이터들을 중첩시켜 볼 수 있는 기능이다. 그림 4 이에 해당하는 메뉴이며, 그림 5 은 이에 따른 결과물이다. 4번은 해당 염색체, 해당 종족에 존재하고 있는 Exon 영역에 RetroElement가 존재하는지에 대한 필터링을 할 수 있는 기능이며, 그림 7 해당한다.

며, B는 Exon & Intron 영역의 시작 점에서 0.5, 1.0, 2.0, 3.0K 앞에 존재하는 즉, Promoter 영역 안에 존재 또는 Promoter에서 Exon & Intron 에 걸쳐서 나타나는 RetroElement들을 보여주고 있는 것이다. B에 나타나는 것중 초록색 영역이 Promoter를 나타내고 있다. RetroScope에서 제공되는 중첩된 시각화는 각 종족, 염색체에 나타는 RetroElement들을 한번에 중첩된 시각화를 제공한다. 이 서비스의 결과물은 그림 5에 해당한다. 그림 5에 보이는 테이블들은 염색체에 나타나는 데이터들의 정보를 나타내고 있다. 테이블들은 각 RetroElement들의 이름, 그리고 이 데이터들이 나타나고 있는 종족, 염색체를 보여주고 있다. 하단부에 보이는 시각화가 바로 중첩된 데이터들을 보여주고 있는 것이다. 각 데이터들 명칭에 따라서 다른 색깔, 다른 포지션을 제공하여 사용자의 편의를 제공한다. 본 서비스는 너무 많은 데이터에 대하여 중첩된 데이터를 제공하면 가독성이 떨어질 것을 고려하여서 3개 미만으로 제한을 두었다. 본 단락에서는 RetroScope에서 제공되는 시각화 서비스에 대하여 소개되었다. RetroScope에서는 기본적으로 Comparative 한 비교를 위해 사용자 편의를 제공하였고, RetroElement가 Exon & Intron 영역에 존재하는지에 대한 조회 및 시각화를 제공하였다. 또한 RetroElement에 대한 중첩된 시각화를 제공하는 서비스도 제공하였다.

4 Conclusion & Future Work

최근 전유전체의 비교를 통하여 인간 특이적인 유전자를 밝혀 내고 있는 실정이다. 또한 인접한 유전자간의 관계를 종간 전유전체적으로 분석하고 있으며, Browser를 통하여 biologist 또는 사용자에게 제공되고 있다. 우리는 이러한 브라우저, 또는 연구를 분석하였으나, 이러한 원인이 되는 DNA 재배열과 RetroElement의 삽입에 대해서는 알 수가 없다. 그리하여 본 논문에서는 RetroElement 데이터를 시각화하는 툴인 RetroScope를 제안하였다. 본 논문에서 제시하였던 RetroScope는 효과적인 서비스를 위하여 Component Ware을 기반으로 구성되었다. Component Ware 환경은 어떠한 환경에서도 서비스를 제공할 수 있다. 본 논문에서 제시한 RetroScope는 Compoenent Ware 환경을 구성하기 위하여 JAVA 언어로 구성하였다. JAVA 언어의 플랫폼 독립적 특성때문에 다양한 시스템 환경을 위한 애플리케이션 제작에는 JAVA는 좋은 언어가 된다. 브라우저에서 서비스, 그리고 데이터를 구성하고 있는 Database와의 통신 및 데이터 엑세스는 PHP로 구성되었다. PHP는 뛰어난 성능, 그리고 다양한 Database 지원하는 인터페이스 를 지원하며, 브라우저에서의 사용자에게 UI도 제공할 수 있다. RetroElement의 유전체 데이터는 My-SQL을 통하여 데이터 베이스를 구성한다. 위와 같은 환경을 바탕으로 RetroScope는 브라우저상에서 사용자의 선택적인 사항에 따른 데이터를 엑세스 하여 JAVA 언어로 시각화를 구성한다. 제공되는 시각화는 기본적인 RetroElement를 Comparative한 환경으로 제공하며, Exon & Intron 영역에 존재하는 RetroElement를 보여주고 있다. 또한 각 영역에 존재하는 RetroElement들에 대하여 중첩된

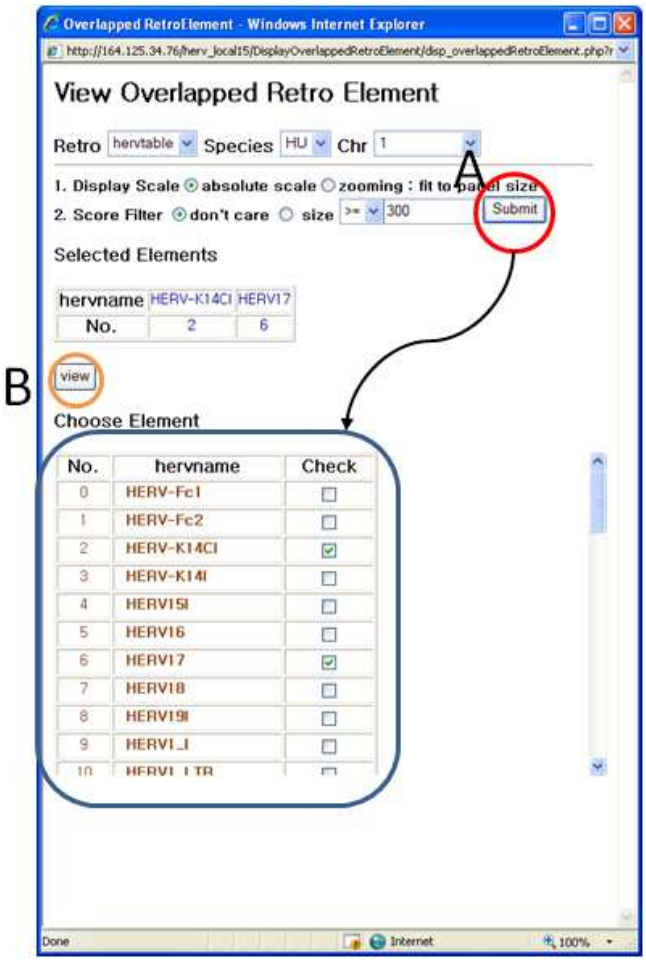


그림 4. 본 그림은 각 염색체에 존재하는 RetroElement를 3개 미만으로 선택하여 중첩된 시각화를 제공할 수 있는 메뉴를 보여주고 있다. A에 해당하는 Submit 버튼을 활성화 할경우 해당 종족, 염색체를 기반으로 하여 존재하는 RetroElement 들이 테이블에 제공된다. 제공된 RetroElement 데이터 들을 체크 박스를 통하여 B에 있는 View 버튼을 활성화 할경우 그림 5와 같은 결과물이 나온다.

시각화를 제공하기도 한다. 본 논문에서 제공된 RetroScope는 RetroElement에 대하여 브라우저를 구성하고 시각화를 제공하였다. RetroElement에 대하여 다양한 시각화를 제공하였다. RetroScope의 향후과제로써는 Exon & Intron 영역에 위치하는 RetroElement들에 대하여 LOD 시각화를 제공하고, 중첩된 RetroElement 들에 대하여 유사도를 평가하는 것이다. 유사도를 평가하기 위하여 Alignment을 통하여 유사도를 계산하여야 한다.

참고 문헌

1. Ewan Birney, T. Daniel Andrews, Paul Bevan, Mario Caccamo, Yuan Chen, Laura Clarke, Guy Coates, James Cuff, Val Curwen, Tim Cutts, Thomas Down, Eduardo Eyras, Xose M. Fernandez-Suarez, Paul Gane, Brian Gibbins, James Gilbert, Martin Hammond, Hans-Rudolf Hotz, Vivek Iyer, Kerstin Jekosch, Andreas Kahari, Arek Kasprzyk, Damian Keefe, Stephen Keenan, Heikki Lehvaslaiho, Graham McVicker, Craig Melsopp, Patrick Meidl, Emmanuel Mongin, Roger Pettett, Simon Potter, Glenn Proctor, Mark Rae, Steve Searle, Guy Slater, Damian Smedley, James Smith, Will Spooner, Arne Stabenau, James Stalker, Roy Storey, Abel Ureta-Vidal, K. Cara Woodwark, Graham Cameron, Richard Durbin, Anthony Cox, Tim Hubbard, and Michele Clamp, An Overview of Ensembl, *Genome research* **14** (2004), 925–928.

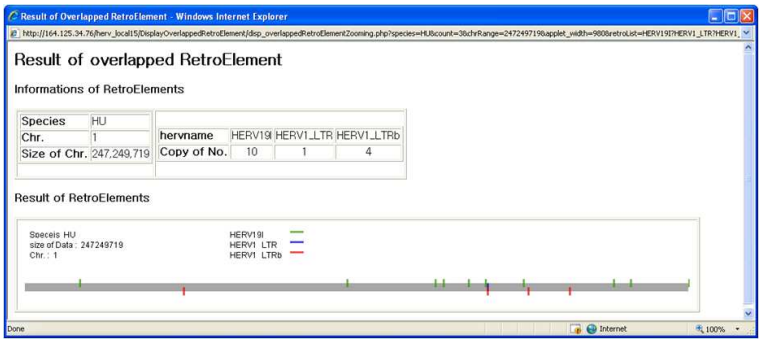


그림 5. 본 그림은 각 염색체에 존재하는 RetroElement의 중첩된 결과를 보여주고 있다. 3개 미만으로 선택된 RetroElement들의 정보들은 테이블로 보여주고 있으며, 시각화를 제공하는 회색바는 전체 영역이며, 3가지의 색갈로 구분지어서 데이터들을 나타내고 있다.

표 1. Table1

Species	Chr	Size	No of Copies
HU	3	199501827	9
RH	3	196418989	14
CH	10	135001995	9
CH	6	173908612	12
OR	12	136387465	10

2. Kushal Chakrabarti and Lior Pachter, Visualization of Multiple Genome Annotations and Alignments With the K-BROWSER, *Genome research* **14** (2004), 716–720.
3. Andrew B. Conley, Jittima Piriyapongsa, and I. King Jordan, Retroviral promoters in the human genome, *Bioinformatics* **24** (2008), 1563–1567.
4. Robert D. Finn, James W. Stalker, David K. Jackson, Eugene Kulesha, Jody Clements, and Roger Pettett, ProServer, *Bioinformatics* **23** (2007), 1568–1570.
5. Jeffrey Heer, Stuart K. Card, and James A. Landay, prefuse: a toolkit for interactive information visualization, CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems (New York, NY, USA), ACM, 2005, pp. 421–430.
6. Gregg A. Helt, Suzanna Lewis, Ann E. Loraine, and Gerald M. Rubin, BioViews: Java-Based Tools for Genomic Data visualization, *Genome research* **8** (1998), 291–305.
7. Ela Hunt and Neil Hanlon, SyntenyVista, NordiCHI '04: Proceedings of the third Nordic conference on Human-computer interaction (New York, NY, USA), ACM, 2004, pp. 455–456.
8. Joanna Jakubowska, Ela Hunt, Matthew Chalmers, Martin McBride, and Anna F. Dominiczak, VisGenome, *Bioinformatics* **23** (2007), 2641–2642.
9. W. James Kent, BLAT - The BLAST-Like Alignment Tool, *Genome research* **12** (2002), 656–664.
10. W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Pringle Tom H., Zahler Alan M., Haussler, and David, The Human Genome Browser at UCSC, *Genome research* **12** (2002), 996–1006.
11. Dae-Soo Kim, Chi-Young Cho, Jae-Won Huh, Heui-Soo Kim, and Hwan-Gue Cho, EVOG: a database for evolutionary analysis of overlapping genes, *Nucleic acids res.* **37** (2009), 698–702.
12. SE Lewis, SMJ Searle, N Harris, M Gibson, V Iyer, J Richter, C Wiel, L Bayraktaroglu, E Birney, MA Crosby, JS Kaminker, BB Matthews, SE Prochnik, CD Smith, JL Tupy, GM Rubin, S Misra, CJ Mungall, and ME Clamp, Apollo: a sequence annotation editor, *Genome biology* **3** (2002), research0082.1–0082.14.

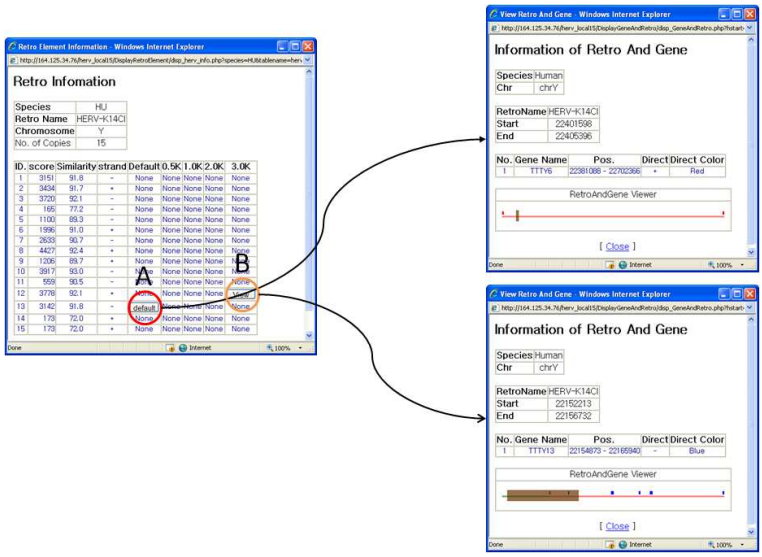


그림 6. 본 그림은 각 영역, 염색체에 존재하는 Exon 영역에 RetroElement가 존재하는지에 대한 여부를 나타내고 있다. default의 경우 Exon 영역안에 RetroElement가 상주 하고 있는 경우이며, 0.5K 의 경우는 Exon 시작 좌표 0.5K 안에 내포하고 있는 경우이다. 1.0, 2.0, 3.0K 이하 동일하다. A의 경우는 일반적인 정보들은 테이블로 나타내고 있으며, Exon 영역은 붉은색으로 표현된다. 방향이 라서 붉은색으로 표현되었다. RetroElement는 갈색으로 표현되어 Exon 영역안에 RetroElement가 상주하고 있음을 알 수 있다. B의 경우 Promoter와 Exon 영역에 넓게 걸쳐서 상주 하고 있는 모습을 나타내고 있다. 녹색선은 Promoter를 나타내며, 파란색은 Exon을 나타내며, 갈색 역시 RetroElement를 나타내고 있다.

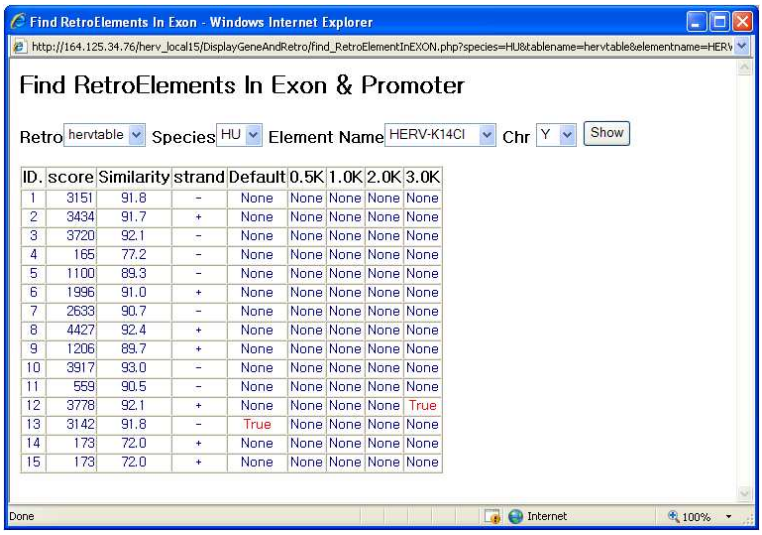
13. Feng Lu, Ji Zhang, and Yanhong Zhou, A Computational Framework and Browser for Supporting Automatic Genome Annotation, GCCW '06: Proceedings of the Fifth International Conference on Grid and Cooperative Computing Workshops (Washington, DC, USA), IEEE Computer Society, 2006, pp. 389–396.

14. David Nix and Michael Eisen, GATA: a graphic alignment tool for comparative sequence analysis, *Bmc bioinformatics* **6** (2005), 9.

15. Kim Rutherford, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice, Marie-Adele Rajandream, and Bart Barrell, Artemis: sequence visualization and annotation, *Bioinformatics* **16** (2000), 944–945.

16. Nameeta Shah, Olivier Couronne, Len A. Pennacchio, Michael Brudno, Serafim Batzoglou, E. Wes Bethel, Edward M. Rubin, Bernd Hamann, and Inna Dubchak, Phylo-VISTA: interactive visualization of multiple DNA sequence alignments, *Bioinformatics* **20** (2004), 636–643.

17. Lincoln D. Stein, Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, Allen Day, Elizabeth Nickerson, Jason E. Stajich, Todd W. Harris, Adrian Arva, and Suzanna Lewis, The Generic Genome Browser: A Building Block for a Model Organism System Database, *Genome research* **12** (2002), 1599–1610.



ID.	score	Similarity	strand	Default	0.5K	1.0K	2.0K	3.0K
1	3151	91.8	-	None	None	None	None	None
2	3434	91.7	+	None	None	None	None	None
3	3720	92.1	-	None	None	None	None	None
4	165	77.2	-	None	None	None	None	None
5	1100	89.3	-	None	None	None	None	None
6	1996	91.0	+	None	None	None	None	None
7	2633	90.7	-	None	None	None	None	None
8	4427	92.4	+	None	None	None	None	None
9	1206	89.7	+	None	None	None	None	None
10	3917	93.0	-	None	None	None	None	None
11	559	90.5	-	None	None	None	None	None
12	3778	92.1	+	None	None	None	None	True
13	3142	91.8	-	True	None	None	None	None
14	173	72.0	+	None	None	None	None	None
15	173	72.0	+	None	None	None	None	None

그림 7. 본 그림은 Exon 영역안에 RetroElement가 존재하는지에 대한 여부를 판단하는 것이다. 그림을 보면 알수 있듯이, 데이터가 존재하는 곳에는 붉은 글씨로 True라고 쓰여져 있으며, 존재하지 않은 경우 None으로 표기하고 있다.