Assessment of genome visualization tools relevant to HIV genome research: Development of a genome browser prototype

AP Boardman

Examination Copy

Keywords

HIV

Data visualization

Genome browser

Genomics

Assessment criteria

HIV databases

Epitope data

GBrowse

Abstract

Over the past two decades of HIV research, effective vaccine candidates have been elusive. Traditionally viral research has been characterized by a gene-by-gene approach, but in the light of the availability of complete genome sequences and the tractable size of the HIV genome (around 9-10 kilobases), a genomic approach may improve insight into the biology and epidemiology of this virus. A genomic approach to finding HIV vaccine candidates can be facilitated by the use of genome sequence visualization. The rationale behind development of an HIV genome browser is supported by the success achieved through the use of genome browsers in the eukaryotic community. Genome browsers have been used extensively by various groups to shed light on the biology and evolution of several organisms including human, mouse, rat, *Drosophila* and *C.elegans*. Application of a genome browser to HIV genomes and related annotations can yield insight into forces that drive evolution, identify highly conserved regions as well as regions that yields a strong immune response in patients, and track mutations that appear over the course of infection. Access to graphical representations of such information is bound to support the search for effective HIV vaccine candidates.

This study aims to answer the question of whether a tool or application exists that can be modified to be used as a platform for development of an HIV visualization application and to assess the viability of such an implementation. Existing applications can only be assessed for their suitability as a basis for development of an HIV genome browser once a well-defined set of assessment criteria has been compiled.

A brief overview of current HIV sequence analysis databases will be given and their visualization tools (or lack thereof) will be discussed. With the previously defined assessment criteria in mind, four available genome browsers will be evaluated for their usefulness in HIV genome visualization. The browsers include UCSC's Genome Browser, Ensembl, the Generic Genome Browser (GBrowse) that forms part of the Generic Model Organism Database Project, and Map Viewer from NCBI. Finally the prototype that resulted from this study will be discussed in the context of some use-cases.

Declaration

I declare that "*Assessment of genome visualization tools relevant to HIV genome research: Development of a genome browser prototype*" is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Anelda Philine Boardman November 2004

Signed: ______________________________

Acknowledgements

Special thanks to my supervisor, Professor Winston Hide, for the support, guidance and motivation he has given me. He introduced me to the fascinating world of Bioinformatics in which I hope to spend some exciting years.

I would also like to thank the SANBI team for their support and technical assistance. I would not have been able to finish this project without the help of Janet Kelso, Annette Badenhorst, Cathal Seoighe, Alan Powell, Zayed Albertyn, Marcus Collins, Nicki Tiffin, and Nicky Mulder.

The CAPRISA staff at the University of Cape Town was always willing to give input. I appreciate their patience and willingness to help. Special thanks to Helba Bredell.

I would like to express my gratitude to the GBrowse development team, especially Lincoln Stein and Scott Cain, for speedy replies on emails to the GBrowse mailing list.

A word of thanks to my friends, especially Corne and Henry Verhoeven, whom were always willing to listen and give advice or provide IT support at home.

Finally I want to express my deepest gratitude towards my parents for their love and support and for always believing in me.

Table of Contents

Abbreviations

3D	Three dimensional
ACeDB	A *C. elegans* Database
AI	Accute Infection
AIDS	Acquired immunodeficiency syndrome
API	Application programming interface
ARV	Antiretrovirals
BLAST	Basic local alignment search tool
C. briggsae	*Caenorhabditis briggsae*
C. elegans	*Caenorhabditis elegans*
CAPRISA	Centre for the AIDS Programme of Research in South Africa
CCR5	C-C chemokine receptor 5
CD	Cluster of differentiation (i.e. CD4)
CDS	Coding sequence
CGI	Common Gateway Interface
CPU	Central processing unit
CTL	Cytotoxic T-lymphocyte
CXCR4	CXC chemokine receptor 4
DAS	Distributed Annotation System
DHHS	Department of Health and Human Services
DNA	Deoxyribonucleic acid
ESKOM	The Electricity Supply Commission
EST	Expressed sequence tag
GB	Gigabyte
GBrowse	Generic Genome Browser
GDE	Genetic Data Evironment
GFF	Gene feature format
GMOD	Generic Model Organism Database
GNU	GNU's not Unix
GUI	Graphical user interface
HGP	Human Genome Project
HIV RDI	Human immunodeficiency Response Database Initiative
HIV	Human immunodeficiency virus
HIV-1	Human immunodeficiency virus type 1
HIVRT & PrDB	Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database
HLA	Human leukocyte antigen

HTML	Hyper text markup language
LANL	Los Alamos National Laboratory
LTR	Long terminal repeat
MB	Megabytes
Mbp	Mega base pairs
MHC	Major histocompatibility complex
NCBI	National Centre for Biotechnology Information
NIAID	National Institute of Allergy and Infectious Diseases
NICD	National Institute of Communicable Diseases
NIH	National Health Institute
PC	Personal computer
PCR	Polymerase chain reaction
PDF	Portable document format
RAM	Random access memory
RNA	Ribonucleic acid
SAAVI	South African AIDS Vaccine Initiative
SGD	*Saccharomyces* Genome Database
SNP	Single nucleotide polymorphism
SQL	Structured query language
STS	Sequence tagged site
tRNA	Transporter-Ribonucleic acid
UCSC	University of California Santa Cruz
UCT	University of Cape Town
UNAIDS	Joint United Nations Programme on HIV/AIDS
URL	Universal resource locate
US	United States
USA FDA	United States of America Food and Drug Administration
WHO	World Health Organization

Chapter 1 Introduction

In recent years human immunodeficiency virus (HIV) has become the cause for great concern with regards to human health, education, agriculture, industry, human resources, and general human welfare – especially in developing countries like South Africa.

Sub-Saharan Africa houses just over 10% of the world's population but the same region accounts for almost two-thirds of all people living with HIV – between 23.1 and 27.9 million people (http://hivinsite.ucsf.edu/global?page=cr09-00-00). According to the UNAIDS update in 2003, in this year alone, between 2.6 million and 3.7 million people in sub-Saharan Africa became newly infected with HIV, and there were between 2.0 and 2.5 million AIDS-related deaths.

Estimates from UNAIDS and the WHO indicated that at the end of 2003 about 5.3 million adults and children are living with HIV/AIDS in South Africa. In a population of over 46 million people that amounts to an estimated 21.5%. According to Statistics South Africa, the accumulated AIDS-related deaths up to 2004 were estimated to be 1.49 million (Lehohla, 2004).

It is clear that a dire need exists for the development of drugs and vaccines for the treatment and prevention of infection by HIV (Klausner et al., 2003). Studies to identify drug targets and vaccine candidates in HIV have been undertaken in several countries including the United States of America (http://www.niaid.nih.gov/daids/therapeutics/research/hivtherapeutics.htm), China (http://www.nih.gov/news/pr/jun2002/niaid-28.htm), Australia and South Africa (http://www.caprisa.org, http://www.saavi.org.za).

1.1. The Human Immunodeficiency Virus (HIV)

HIV is a retrovirus with genetic material consisting of two RNA strands that integrate into the host genome upon infection and subsequent reverse transcription. The structure of an HIV-1 molecule (Figure 1-1) can be divided into two compartments: the viral envelope and the protein core (Abbas & Lichtman, 2001). The protein core is made of p24 protein from the gag gene and contains the dimeric single stranded RNA molecule associated with a lysine3-tRNA, as well as p7 protein (also known as nucleocapsid), and three enzymes known as protease, reverse transcriptase, and integrase (Crandall, 1999). Matrix protein (p17) surrounds the core and is enclosed by the lipid viral envelope which is derived from the membrane of the host cell. A complex viral protein, *env*,is presented on the outside of the viral envelope. The complex consists of a stem (gp41) and a protruding cap (gp120) which is essential for viral entry into the host cell.

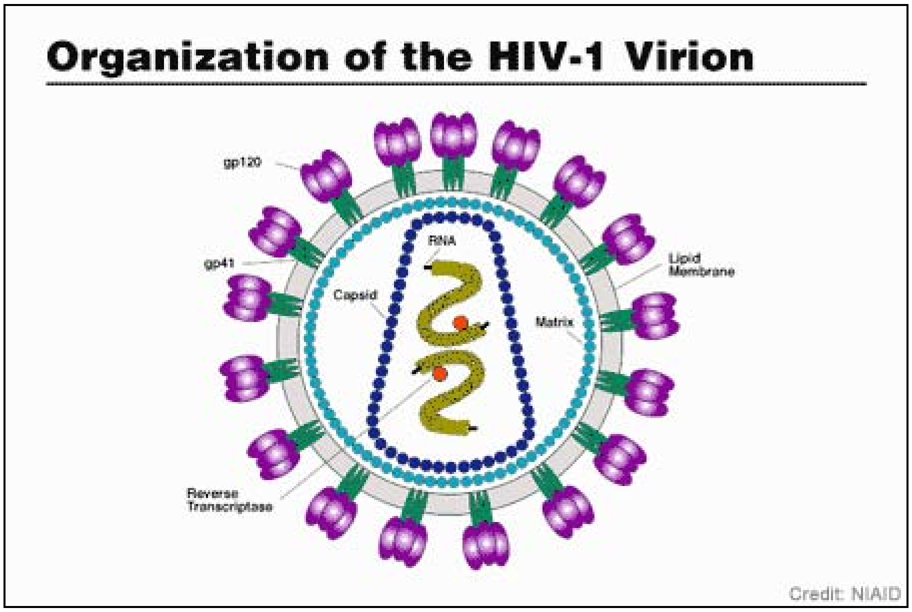


Figure 1-1 The structure of an HIV virion (Copyright: NIAID - http://www.avert.org/pictures/hivstructure.htm)

The HIV genome consists of nine genes flanked by a 5' and 3' untranslated region. The genes can be divided into genes coding for either structural proteins (*gag*, *pol*, and *env*), regulatory proteins (*tat* and *rev*), or accessory proteins (*nef*, *vif*, *vpr*, and *vpu*) (http://hivinsite.ucsf.edu). Upon transcription and translation of genes, resulting peptides are cleaved and modified to yield the 17 known active proteins. Gene products and their functions are described in Table 1-1. Figure 1-2 provides a graphical representation of the HIV genome, showing the nine genes and various other features.

Gene	Protein	Size (amino acids)	Function
Gag	p17 (Matrix protein)	131	• Stabilizes the viral particle.
	p24 (Core-antigen capsid)	231	• Forms the core of the HIV molecule
	p2	14	• Unknown function
	p7 (Nucleo-capsid protein)	55	• Incorporates viral RNA into the HIV molecule
	p1	16	• No known function
	p6	52	• Plays a role in incorporation of Vpr into the assembling virion
Pol	p51 (Reverse Transcriptase)	440	• Forms dimer with p66 that is responsible for reverse transcription of viral RNA
	p15 (RNaseH)	120	• Removes RNA template from newly synthesized DNA to allow synthesis of the complementary DNA strand
	p66 (RT/RNaseH)	560	• Forms dimer with p51 that is responsible for reverse transcription of viral RNA
	p15 (Protease)	99	• Required for cleavage of Gag, Gag-Pol, Pol and Nef precursors
	p31 (Integrase)	288	• Responsible for integration of viral DNA into the host genome

Gene	Protein	Size (amino acids)	Function
Env	gp120 (Surface protein)	481	• Serves as receptor protein on viral envelope – bind to CD4 receptors on host cells
	gp41 (Transmembrane protein)	345	• Mediates fusion of viral and host cell membranes
Rev	p19 (Anti-repression transactivator protein)	116	• Regulates expression of structural and regulatory genes • Transports unspliced RNAs from nucleus to cytoplasm • Responsible for the transition from early to late phase of HIV gene expression
Tat	p14 (Transactivating regulatory protein) (Early, fully spliced RNA)	72	• Viral transcription initiation and/or elongation • Upregulates expression of all viral gene
	p16 (Transactivating regulatory protein) (Late, incompletely spliced RNA)	86	
Nef	p27-p25 (Negative Factor)	123	• Interaction with host cell signal transduction proteins • Down regulation of cell-surface protein like CD4 and MHC Class I expression • Induction of apoptosis in non-infected cells

Gene	Protein	Size (amino acids)	Function
Vif	p23 (Virion Infectivity Factor)	192	• Plays role in viral infectivity
Vpr	p12/p10 (Viral Protein R)	78	• Restrain cell division • Facilitates nuclear localization of preintegration complex
Vpu	p16 (Viral Protein U)	81	• Downregulation of CD4 • Promotes virion release from the plasma membrane

Table 1-1 The HIV genome consists of nine genes resulting in 17 protein products. The table was adapted from information obtained at http://www.bioafrica.net/proteomics/.

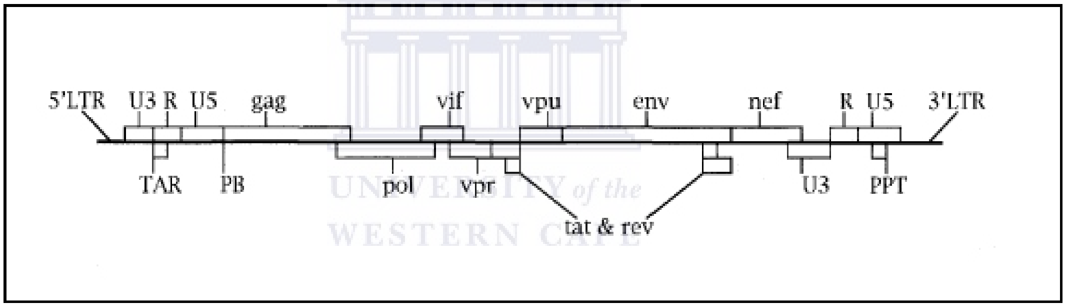


Figure 1-2 A graphical representation of features present on the HIV genome (T. Budd, http://www.it.stlawu.edu/~tbudd/hivgenome.html).

HIV infects a variety of immune cells – mainly CD4+ T lymphocytes and macrophages (Pomerantz, 2003). The life cycle of HIV consists of sequential steps briefly outlined below (Abbas & Lichtman, 2001):

(a) Infection of host cells

(b) Reverse transcription of viral RNA into double stranded DNA

(c) Integration of viral DNA into the genome of the host cell

(d) Expression of viral genes

(e) Production of new virions

1.2. HIV Research in South Africa

Two research projects that are directly linked with the HIV Genome Browser study are currently underway in South Africa. In 2002 a large collaborative project aimed at understanding the pathogenesis and epidemiology of HIV, and using this knowledge to develop HIV/AIDS treatment and prevention, was formed in South Africa. The Center for the Aids Programme of Research in South Africa (CAPRISA) (http://www.caprisa.org) is funded by the US National Institute of Health (NIH), National Institute of Allergy and Infectious Diseases (NIAID), and the US Department of Health and Human Services (DHHS). Institutes collaborating on the CAPRISA project in South Africa include the University of Natal, University of the Western Cape, University of Cape Town, and the National Institute of Communicable Diseases (NICD).

CAPRISA consists of a number of different projects including the Acute Infection (AI) and Data Integration projects. The AI project studies viral set point and clinical progression in HIV-1 subtype C infection. In this study, the role of immunological and viral factors during acute and early HIV infection is under investigation (http://www.caprisa.org/Projects/acute_infection.html). The Data Integration project aims to develop and maintain databases for the integration of molecular, clinical and immunological data generated by studies like the AI project and to facilitate data sharing between various participating institutes. The systems that will be responsible for achieving these aims are currently under development. Table 1-2 illustrates the variety of data types that will be supplied by institutes participating in the CAPRISA project.

Institution	Data-types
University of Natal	• Patient information (patient ID, age, sex, geographic location, etc) • Clinical data for patient (date of last virus negative, clinical symptoms, etc) • Behavioural data
National Institute of Communicable Diseases	• CD4 and CD8 count • Viral load (p24 isolation) • CTL response • Magnitude and breadth of neutralizing antibody response • HLA type and HLA restriction information • Viral phenotype: CCR5, CXCR4 • Source material (plasma, DNA, serum, other) • CTL reactive peptide or epitope information
University of the Western Cape	• Responsible for data integration and data management • No data generated
University of Cape Town	• Complete genome sequences (to be sequenced by Stellenbosch University) • Genome fragment sequences (to be sequenced by Stellenbosch University)

Table 1-2 A summary of institutions forming part of the CAPRISA project and the relevant data types that will be generated by each. (CTL Cytotoxic T-lymphocyte, HLA Human Leukocyte Antigen)

The South African AIDS Vaccine Initiative (SAAVI – http://www.saavi.org.za) was established in 1999 and is responsible for research, development and testing of HIV/AIDS vaccines in South Africa. SAAVI is a lead programme of the Medical Research Council of South Africa and is primarily being funded by the Department of Health, Department of Science and Technology and ESKOM.

Through projects executed by CAPRISA and other HIV research programmes, an array of data types will be made available to the HIV research community of South Africa. These data types include molecular, clinical, behavioural, and demographic data (refer Table 1-2). In order to extract as much information as possible from the relevant data, a proposal was made to design a genome browser where molecular and immunological entities can be viewed graphically in relation to each other and relative to a well-characterized reference genome. The browser would serve as one of a selection of entry points into an underlying integration database containing all relevant information. Graphical display of data enables the researcher to perform visual manual data mining since it is possible to easily identify interesting genomic regions at once.

1.3. The History of Genome Browsers

The Human Genome Project (HGP) formally began in October 1990 (Choudhuri, 2003). The HGP is an international collaboration between laboratories from the United States, the United Kingdom, Germany, France, Japan, and China. The aim of the HGP was to have a completely sequenced human genome by 2005. In the process of sequencing the human genome, it became clear that sequencing other organisms would support this project directly as well as indirectly. Bacterial genomes were sequenced with the hope of shedding light on the molecular basis of pathogenesis. *Caenhorhabditis elegans'* genome could facilitate understanding of development. *Drosophila* is an organism that has been studied extensively - the genome sequence would add to the array of tools and data available to the fruitfly research community. Sequencing other genomes could also help to solve problems encountered in the sequencing of the human genome. In 1995 the 1.8-Mbp genome of *Haemophilus influenzae* became the first genome to be sequenced completely (Fleischmann et al, 1995). Since then 222 complete genome sequences have been published. Currently 1199 genome projects are described at the Genomes OnLine Database Web site, including 522 prokaryotic and 453 eukaryotic genomes (http://www.genomesonline.org/).

With the accumulation of sequence data, researchers needed tools to store, display, and contextualize their data. The earliest form of Genome Browser was probably released in June 1991 with the development of ACeDB (A *Caenorhabditis elegans* Database) (Stein &

Thierry-Mieg, 1998). ACeDB was developed to manage and distribute genetic data of *C. elegans* (Stein & Thierry-Mieg, 1999). ACeDB is an open source project that provides a database management system together with methods for managing DNA and protein sequences, genome maps, and other entities that could be encountered. The standalone application provides graphical displays for a wide range of specialized biological data. Since the development of ACeDB for the nematode research community, it has been used in various other organism databases such as the human sequencing project at the Sanger Center and the Washington University Genome Sequencing Center (Stein & Thierry-Mieg, 1998), Xenopus laevis - Axeldb (Pollet et al. 2000), and mycobacterium (Bergh & Cole, 1994).

The next step was to make genome browsers more accessible to the research community. The Saccharomyces Genome Database was designed to be accessible through the World Wide Web to potential users (Cherry et al. 1998).

Developers became aware of needs voiced by the user community and started developing tools to deal with these requirements (Loraine & Helt, 2002). User requirements include the ability to:

a) view the sequence of their organism of interest at different levels of resolution
b) track clones and contigs
c) view chromosomes and banding patterns (in the case of the human genome)
d) download raw sequence data for further analyses

e) view annotations associated with the raw sequence (Annotation examples include known genes, ESTs, predicted genes, introns, STS markers, SNPs and many more)

f) have links from sequence data to more information regarding specific features (examples would include information regarding functions of proteins, metabolic pathways implicated, gene expression, known mutations, and associated diseases)

g) search the genome via accession numbers, similarity searches or keywords

A genome browser is thus a database of sequences, annotations, and links to external sites that is interfaced by a graphical user interface (GUI) (and oftentimes a web-based GUI). The GUI allows a user to browse the genomes in a simple and efficient manner. According to the genomics glossary of the Cambridge Healthtech Institute "The genome browser itself does not draw conclusions; rather, it collates all relevant information in one location, leaving the exploration and interpretation to the user." (http://www.genomicglossaries.com/content/genomics_glossary.asp).

1.4. HIV and Genome Browsers

Historically viral research has followed a gene-by-gene approach instead of the high-throughput genomic approach that has been taken for eukaryotic and bacterial genome studies. Sequence-centric visualization applications for eukaryotic genome annotations play an important role in eukaryotic genome analysis and discovery. The Ensembl genome viewer (http://www.ensembl.org) allows gene structure analysis and comparison of support obtained from predictions of novel genes. It also facilitates comparison of different genomes from diverse organisms in order to identify regions of synteny. Other examples of genome database projects that provide genome viewers include NCBI (Map Viewer at http://www.ncbi.nlm.gov/mapview), the Golden Path Assembly of the human genome at the University of California at Santa Cruz (http://genome.ucsc.edu), Wormbase (http://www.wormbase.org), Flybase (http://www.flybase.org), and Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccaromyces). The success that has been achieved using genome browsers to understand the complex biology of eukaryotes like *Homo sapiens*, *Drosophila* and *C. elegans* serves as a motivation for the development of an HIV genome browser, although viral genomes are generally much smaller than eukaryotic genomes and most viral genomes are generally well-annotated in terms of confirmed genes. The types of questions asked in HIV research may differ from those asked in eukaryotic research but genome browsers may be usefully employed to address such questions.

Different HIV databases exist – examples include the Los Alamos HIV databases at http://www.hiv.lanl.gov and Stanford HIV Drug Resistance Database at http://hivdb.stanford.edu - yet visualization tools that match the functionality of those available for analysis of eukaryotic genomes are not yet available for the HIV research community. Developing tools that can facilitate genome analysis through visualization contributes to the process of understanding HIV biology and epidemiology and may well lead to enhanced discovery of vaccines. In South Africa HIV sequence data is being generated by the CAPRISA project which provides the opportunity to be at the forefront of the development of visual analysis tools.

It is important to note that a genome browser-type application is not capable of supporting the whole range of questions that could possibly be asked by researchers. Instead it contributes to the repertoire of tools available to analyse and understand the complexity of genomes. The availability of a genome browser does not remove the need for a data warehouse or other database that could support textual data mining or general SQL-like queries. A genome browser could, however, focus a researcher's attention on visibly interesting areas in the genome of an organism under investigation.

In the development of an HIV genome browser, one of two approaches can be followed: (a) a novel tool can be developed to achieve predefined functionality; or (b) existing software can be evaluated for their suitability to be used as a platform for the development of an HIV-specific genome browser. In this study established genome browsers were evaluated for their applicability to be used as a basis for an HIV Genome Browser and a prototype was susequently developed. A list of requirements for development of an effective HIV browser is discussed in Chapter 2.

Chapter 2 Defining Assessment Criteria of a System Suitable for HIV Genome Visualization and Discovery

Before the developer can design a genome browser for HIV it is essential to understand the biological requirements and technical limitations. The developer has to understand the projects of participating researchers from various disciplines (e.g. molecular biology, epidemiology, immunology) who will use the system to visualize their data (Chen & Carlis, 2002). It is the task of the developer (in collaboration with the end-users) to envision possible features that the browser will have to provide in later stages of the project. Defining assessment criteria for a successful and efficient genome browser and analysis system is a non-trivial problem.

2.1. Biological Requirements

2.1.1. Identifying Questions That Should be Addressed by the Browser

Before currently available software can be assessed for suitability as a basis for development of something more specialized, it is important to define why and how the end product will be employed to support a researcher's work. A broad range of research questions can be addressed through the use of a genome browser depending on the type of genomes and annotations available and the aims of the researcher who will be using it.

Throughout the development process high levels of interaction with future users of the system was maintained in order to define possible questions that the browser should be able to address. Interviews were conducted with various researchers and principle investigators participating in the AI study as well as the Data Integration project. Interviews included discussion about each scientist's research question (or questions) and the datatypes generated by or required for their projects. An intensive study of CAPRISA project documentation was performed in order to have a more complete picture of the type of questions that are asked by HIV researchers. A general literature review of current HIV studies around the world showed that the questions being asked by South African scientists are globally addressed problems – a browser that can address local questions will certainly also be useful to the international scientific community.

The HIV genome browser should for instance be able to facilitate assessment of viral diversity in a specified set of sequences. Studying the level of diversity of the HIV genome is one of the major objectives of HIV research and can highlight regions that are more prone to mutation or that are highly conserved.

Epitope maps display the distribution of short peptides (that elicit immune reactions in the human body) across HIV-1 gene products. Through the availability of epitope maps, researchers can identify regions that elicit a good immune response and ones that have not been proven to cause any immune reaction. Investigators also want to be able to relate certain epitope responses to the HLA type of patients.

2.1.2. Defining Data Types

Questions that can be addressed through the browser are confined by data and relationships between data-types stored in the database supporting the browser. The browser must therefore be developed with a good understanding of the data content of the supporting database. Data types that will be generated in the CAPRISA project were identified through personal communication with the researchers from the various participating laboratories. The data types that were identified are similar to data types being generated by other local and foreign HIV studies.

(i) **Reference genome**

HXB2 (GenBank accession K03455) is generally used as the reference genomic sequence in HIV-1 research. The HXB2 sequence has been annotated extensively with information on coding regions, control elements, epitope maps and other biologically interesting data. Extensive epitope maps relative to the HXB2 genome are available from the Los Alamos Molecular Immunology database. It is therefore important to supply the user with a display of the HXB2 genome and its relevant annotations to use as a reference for their own research.

(ii) **Genomes and genome fragments**

The genome browser has to display complete genomes as well as genomic fragments with reference to a specified genome, for example the HXB2 reference genome of HIV-1 or a genome (or fragment) sequenced at a different time point in the relevant project. The user must be able to visualize his/her proprietary sequences relative to the reference genome even if these sequences have not yet been submitted to the underlying database of the browser.

(iii) Multiple and/or pairwise alignments

Ideally a user will have access to multiple alignments of the data that are displayed in the browser at a specific timepoint. Alignments should be made available for download and use in other analyses. Both nucleotide and amino acid alignments should be made available.

(iv) Proprietary epitope/reactive peptide maps

Reactive peptide and epitope data available from the database that supports the genome browser should be displayed against relevant genome sequences or perhaps consensus sequences. Displaying epitope density over a specified region in the shape of a histogram or graph is very useful to obtain information about regions under strict immunological selection pressure (Ernsthoff, 2002). Ideally the amino acid sequence should be made available to the user and residues that have changed due to mutation in the coding sequence should be indicated.

Epitopes contain biologically significant amino acids as well as neutral residues. Anchor residues are those amino acids involved in binding the HLA molecule and flag residues point towards the T cell receptor (McMichael et al. 2002). Highlighting such residues in the visual display of epitope maps will facilitate easy identification of important regions within the peptides.

(v) Restriction enzyme cut sites

Restriction enzymes are used widely in molecular techniques to generate fragments of nucleic acids on which analyses can be performed. In order to plan future experiments the researcher needs to know at which positions different restriction enzymes will digest the DNA region of interest. If such data can be displayed through the genome browser, the user will not have to use additional external software to determine potential restriction sites in the sequence under investigation.

(vi) Entropy over a selected region for a selected set of sequences

Entropy is calculated to measure amino acid variability at a given codon position. It is calculated for each column in a multiple protein alignment. A larger entropy value is an indication of a higher degree of sequence variation. It has previously been found that higher levels of entropy co-occur with lower epitope density (Ernsthoff, 2002).

(vii) Positive selection over a specified region for selected set of sequences

Positive selection is an evolutionary mechanism whereby newly arisen mutants have higher fitness than the average population, resulting in frequencies of the mutants in the populations increasing over time (Suzuki & Gojobori, 1999). Positive selection plays an important role in the evolution of HIV sequences present in a population over the course of infection and is mainly responsible for the evolution of drug resistance of certain HIV strains (Frost et al, 2001). A low level of selection and high epitope density in a region may be an indicative of a good vaccine candidate (Ernsthoff, 2002).

(viii) Viral Phenotype

HIV uses mainly CCR5 or CXCR4 as co-receptors to bind and enter host cells (Dittmar et al., 1997). The use of a specific co-receptor is associated with the viral phenotype. Users will want to know which co-receptor a specific isolate used to enter and infect immune cells.

(ix) HLA type

Certain HLA alleles have been associated with restriction of HIV infection and/or progress (Moore et al, 2002). HLA typing of patients will be done and made available to researchers in the CAPRISA study. Availability of HLA-allele information may enable researchers to identify interesting patterns of restriction.

2.2. Technical Requirements

2.2.1. General Software Requirements

By means of a literature review on existing Web-based and stand-alone visualization applications aimed at displaying biological data from different sources, it was possible to outline the most important technical requirements of such an *in silico* discovery system. A distinction is made between requirements from the user's point of view and that of the developer's. It is also acknowledged that not all requirements can be met simultaneously but attention will be given to the points that seemed important from the literature review. Chapter 4 explains which of the requirements are met by the prototype.

(i) User requirements

During the design of any software application it is essential to bear in mind the user's computer literacy level. In the case of a genome browser for HIV, the expected users range from bench biologists in small research laboratories who are used to point-and-click GUI applications to bioinformaticians with more knowledge of command line operating systems.

For a useful application to be accepted into the research community it has to be user-friendly. It should preferably have an intuitive user-interface through which the researcher can navigate easily without having to spend a large amount of time learning how to use it (Birney et al. 2002). It should also be well documented so that users have quick access to help and don't have to rely on single persons to assist with technical or other support.

The HIV genome browser is aimed at researchers who already have access to computers as well as the Internet within their laboratories but who may not have a systems administrator to assist with software installations and upgrades. Making the browser available as a Web-based application will thus provide the user with the latest version of the software run on the server through which the browser is accessed (Matthiessen, 2002).

It is worth considering the cost of obtaining a genome browser. Costs can be minimized by providing an Internet-based application developed on the principles of freeware and open-source software.

Using freeware and open-source software makes projects possible through the collaboration of institutes and individuals that would have otherwise not been achievable. Through the principles of the open-source software community reliable, secure software can be developed rapidly and submitted for peer review to ensure the software performs adequately (http://www.opensource.org).

Investigators need access to the most current information available (Dennis et al. 2003, Helt et al.1998). A centralized data repository accessible through a web-based application eliminates the issue of downloading large databases. Access speed is, however, dependent on the number of simultaneous users. Design and implementation of the underlying database is a very important aspect and although it falls outside the scope of this study, it should be taken into account when undertaking design and implementation decisions for the genome browser.

The developed application should not be platform dependent (Stein et al. 1998). By developing a Web-based application developers can build the system on their platform of choice and only have to ensure that it inter-operates with most Web-browsers or at least on Web-browsers that are freely available. In the latter case users could be persuaded to download and install a freely available Web-browser and use the application in this framework.

Another essential aspect that should be considered is the security risk of storing all data on a publicly accessible machine. Researchers might be skeptical of using Web-based applications to display proprietary data for fear that the wrong people might intercept it. It is thus of utmost importance to ensure safe data transmission and display (Laird et al, 2003). There are a number of protocols in use to ensure the safety of data. These will not be discussed here since they form an integral part of the Web server and Web browser (outside the scope of this project).

According to Helt et al (1998) users of GUIs (graphical user interface) respond better to an interface that is responsive and interactive. Response times upon interaction with the system should be shorter than 1 second - longer response times (over 15 seconds) are generally detrimental to productivity (Shneiderman, 1984) . Achieving real-time responses can be difficult with a Web-based application, especially if the application supports a large amount of detailed graphics or large volumes of data.

(ii) Developer requirements

One of the advantages of developing Web-based applications is that it can be maintained centrally. Bugs identified in software that is maintained centrally can easily be fixed and new releases of software can be made as often as required. Users should be able to report bugs to the development team and fixes should be made available as soon as possible. Developers can also incorporate general requests for new functionality into the system and have it made available. Web based applications is easier to deploy than standalone applications (A Powell, personal communication).

There are a number of features of existing applications that can be used in the development of new applications. Modular applications are easier to extend or modify to incorporate additional functionality (Robinson & Flores, 1997). Component-based systems permit a developer to customize the software to suit specific needs by only having to change certain components and adding features by adding new components rather than having to alter a huge script. Software should be portable and extendable to make it easier for the developer to adapt it for application in a new area of research (Mungall et al. 2002).

Making use of standardized protocols and methods is another way of ensuring that a system is portable and extendable. If the original developers made use of a language that is commonly used in Bioinformatics application development and employed standard protocols and methods, it reduces the amount of effort and frustration involved in extending the functionality of a software system.

Extending and customizing an existing system can be facilitated by good documentation of the system specification and installation. It is of benefit to the user if developers of the original software are available to give support either directly or through the availability of a helpdesk, mailing list or discussion forum. Good support can usually be obtained from open-source projects. Popular open source projects often provide good support via mailing lists, discussion for a and email where users can get help or information on bug-fixes, new releases and conferences or meetings regarding the software. Feedback on email requests and enquiries is generally speedy.

2.2.2. Genome Browser Specific Requirements

The dataset that will be displayed in the HIV genome browser is quite large and may grow over time to become very difficult to display on a computer screen without causing screen clutter (Chi et al, 1996). Layout and display of the data should support analysis. Glyphs representing different data types play a major role in comprehension of the display. According to the IBM Corporation's glossary a glyph can be defined as "a graphical symbol whose appearance conveys information (http://www.Inf.infn.it/computing/doc/aixcxx/html/glossary/g.htm). Glyphs must be chosen after careful consideration in order to enhance data comprehension. Colour can also be applied to display certain features of the data. Varying the hue of a specific colour has, for example, been implemented to indicate the percentage similarity between two sequences (Stein et al, 2002).

The genome browser must be able to display data accurately (Globus & Uselton, 1995). The "focus+context" paradigm, as described by Robinson and Flores (1997), implies that a user should, even when he/she is zoomed into a small region of his/her data, be able to retain a sense of the context of the area of interest within the bigger picture of the complete genome or chromosome. Similarly, for example the ability of viewing the same data at different levels of detail – known as semantic zooming – is aan effective method of enhancing data comprehension. Semantic zooming refers to viewing the same data at different levels of detail depending on the level of zooming that is chosen. For instance, an expressed sequence tag (EST) alignment to a region in the genome can be displayed as a box at low resolution; at higher resolution the aligned DNA sequences can be displayed.

The database underlying the genome browser, must be able to meet the query needs of individual investigators. Individual researchers want to be able to customize the display to show only data types that are applicable to their research, in order to reduce screen clutter (Navarro et al. 2003).

The genome browser should ideally use a regularly updated database containing all relevant information. A useful feature would be support of the DAS (Distributed Annotation System) protocol as described by Dowell et al (2001). DAS allows users to visualize annotations on the genome of interest made available by laboratories using DAS servers.

Investigators want to be able to see data that has not yet been submitted to the database but resides in a file on their local machine. Such data may never be submitted to the database because it originated from a previous or separate study, but may nevertheless be interesting to view in relation to what is available in the database.

Links to more information regarding the areas of interest should be provided. Clicking on a certain gene can, for example, display a page where the gene, its products, coordinates, and other interesting characteristics are available. Links can also be supplied to related databases or Websites that provide more specialized information. For example if the user clicks on a certain epitope he/she could navigate to the LANL immunological database.

Once the user has identified a region of interest, has zoomed into this particular region via the genome browser and has all the relevant annotations/data-types that are present in this area displayed, it should be possible to download the information to do further analyses. A researcher might, for example, want to edit a multiple sequence alignment or add a sequence to a multiple sequence alignment. It must thus be possible to save the displayed data as a file on the investigator's computer in order for him/her to import such a file into external applications.

Another very important criterion that the ideal browser should satisfy is the support of publishable views. Users want to have access to high quality images that can be incorporated into their articles for publication in scientific journals.

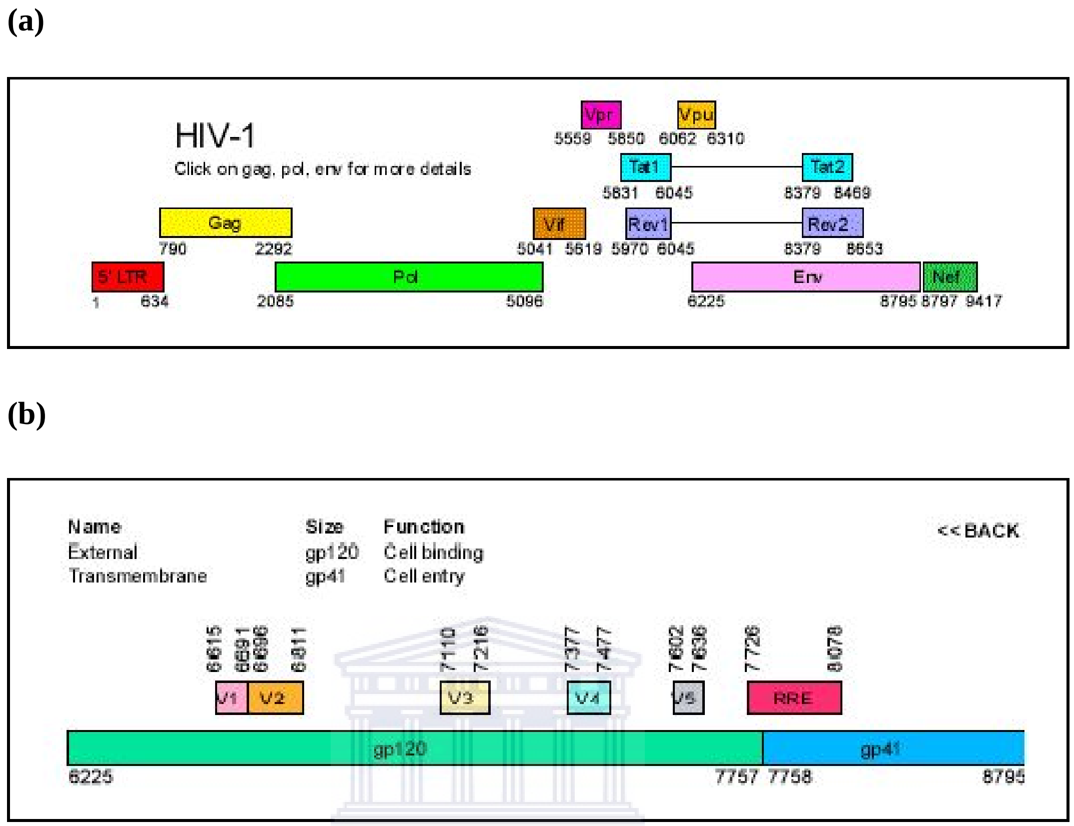
Chapter 3 Critical Comparison of Existing HIV Resources and Available Genome Browsers

3.1. Assessment of Available HIV Databases and Relevant Visualization Tools

3.1.1. Los Alamos National Laboratory (LANL)

The HIV Genetic Sequence, Immunology, and Drug Resistance Databases developed and maintained by the Los Alamos National Laboratory – LANL – (http://www.hiv.lanl.gov) are the most well known HIV related databases amongst researchers in this field. Since 1987 HIV data has been collected, curated and annotated through this project. The Web-based service provides analysis tools, reference alignments, and an extensive collection of HIV and SIV sequences and annotations to researchers.

Sequences are obtained from GenBank and annotated with information from the literature and to a lesser extent information from authors. The database can be accessed through a search interface. Tools provided by LANL include phylogenetic analysis tools, tools that support similarity searches, coordinate mappers on HXB2 and more. In total the HIV Sequence Database currently offers 11 analysis tools and links out to four additional external tools including sequence format conversion, HIV subtyping using BLAST and a restriction site mapper.



Name	Size	Function
External	gp120	Cell binding
Transmembrane	gp41	Cell entry

Figure 3-1 A map of the HIV genome available on the Los Alamos HIV Website. (a) represents the complete genome with all coding genes and the 5' LTR. The coordinates are relative to the HXB2 reference sequence. An expanded view of a region can be displayed by selecting the region. (b) is the expanded view of the *env* gene, detailing the components, names and functions of different subenomic regions.

Despite of all the tools offered at Los Alamos for the HIV researcher, a way to visualize propriatory data is not available. They offer a clickable map of the annotated HXB2 reference genome sequence (available at http://www.hiv.lanl.gob/content/hiv-db/MAP/hivmap.html) representing all 9 genes encoded by the HIV genome, the 5'LTR region and the associated coordinates. If a specific gene is clicked an expanded view giving more information about the particular gene, its sub-sequences and other annotations are displayed (Figure 3-1 a and b).

The available genome map is restricted to HXB2. Newly generated sequences can’t be viewed in their genomic context or referenced to HXB2.

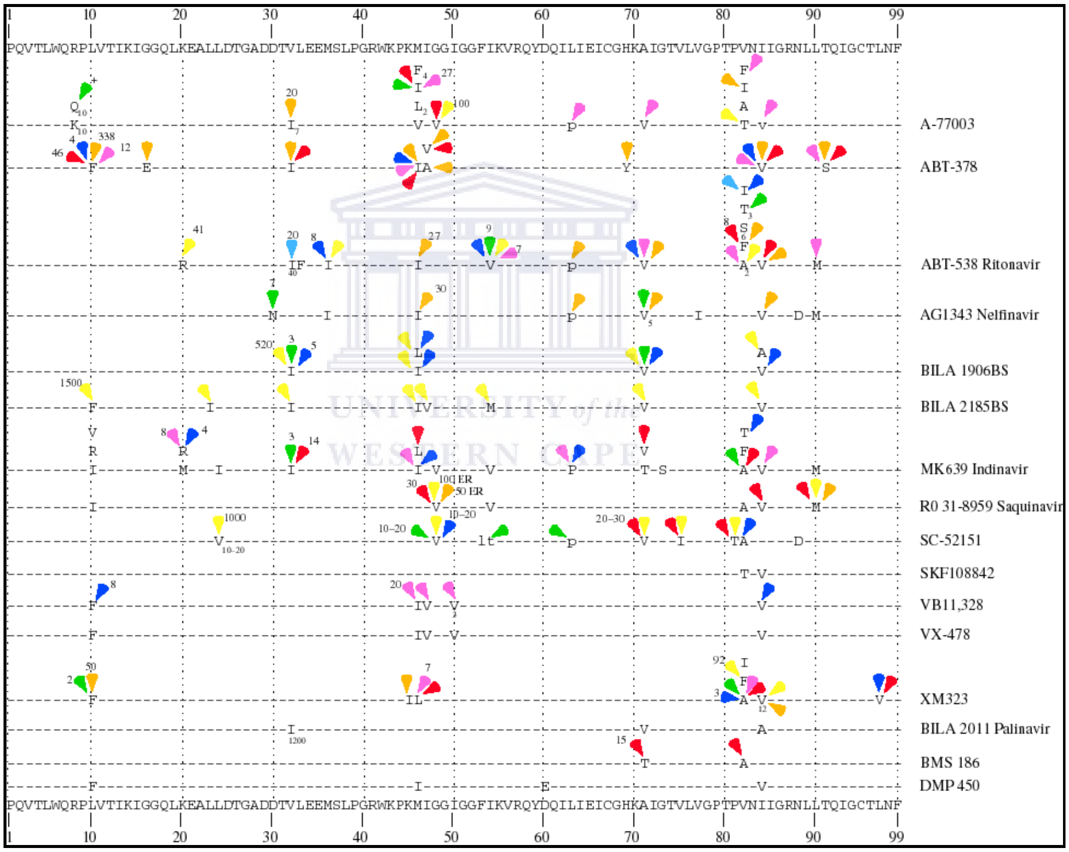


Figure 3-2 A mutation map of the HIV protease gene compiled by the Los Alamos HIV team.

A mutation map is provided through the LANL HIV Drug Resistance Database (Figure 3-2). It is still in a preliminary phase of development and is currently only available for the protease amino acid sequence. In the map, rows represent antiretrovirals and each resistance conferring mutation is indicated by a coloured glyph if it acts in concert with another mutation to confer resistance. A researcher can obtain information about the level of resistance conferred by different mutations. HXB2 is used as the wild-type sequence and every position that correlates to the amino acid found in HXB2 is indicated with a dash.

The mutation map isn't interactive and the researcher cannot add proprietary observations to the map to view it in context with data that is already available.

From the Drug Resistance Database (http://resdb.lanl.gov/Resist_DB/default.htm) the user has the option of viewing the three-dimensional structure of either the protease or reverse transcriptase proteins. A labeled version of the 3D structure indicates mutations in HIV-1 protease that confer drug resistance. The active site is clearly indicated (Figure 3-3). The user can select regions in the protein that should be highlighted.

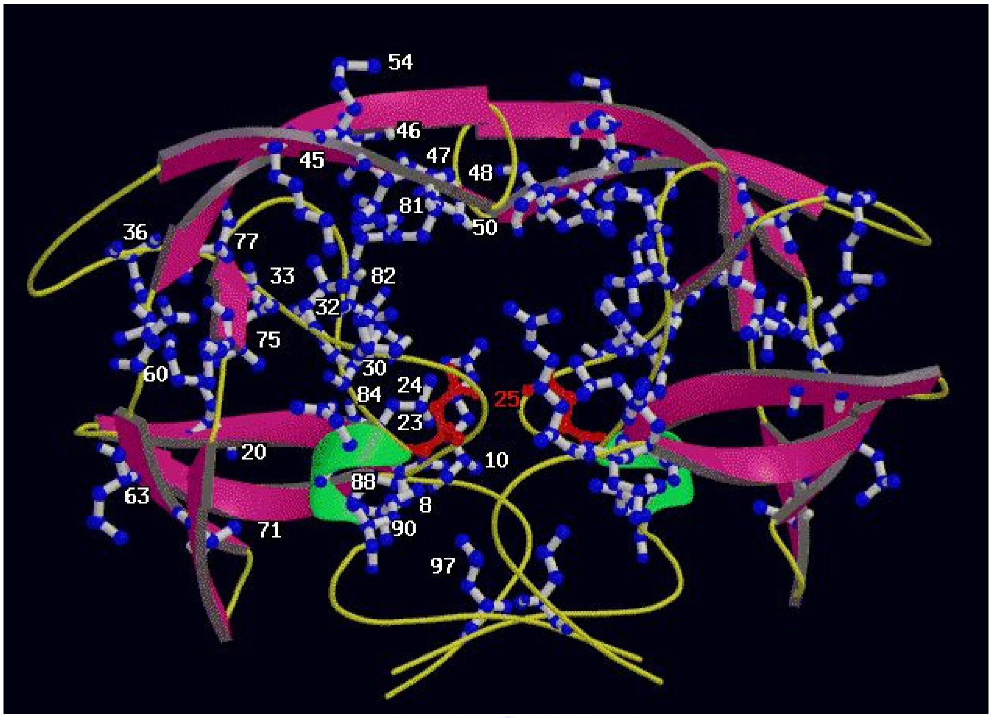


Figure 3-3 A 3D image of HIV protease indexed to indicate important residues (obtained from http://resdb.lanl.gov/Resist_DB/default.htm). The active site is coloured red.

Epitope maps can be obtained from the HIV Molecular Immunology Database (http://hiv.lanl.gov/content/immunology/index). The epitope maps are available in two formats: (1) three static PDF documents comprising between 25 and 33 pages that contain complete maps for CTL (cytotoxic lymphocyte), T-helper cell epitopes as well as antibody epitopes; and (2) HTML versions of the same epitopes mapped to a specific gene or subgenomic regions available at separate URLs. Although all known epitopes are incorporated into these maps, it is difficult to comprehend the true context of the epitopes as they cannot be visualized with reference to the complete genome or other annotations. Figure 3-4 is an extract from the LANL immunology database containing the map of known epitopes in the Rev protein.

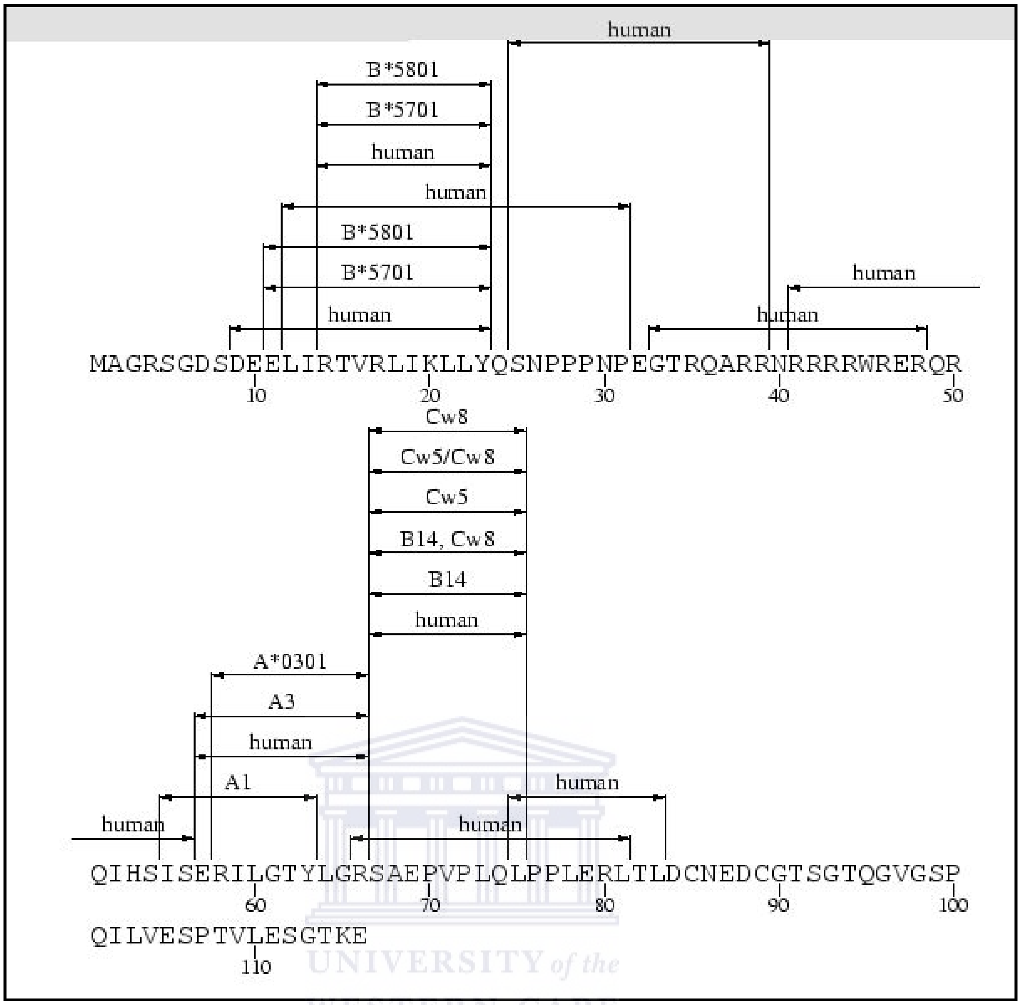


Figure 3-4 A screenshot of the CTL epitope map available for the HXB2 Rev protein sequence from the Los Alamos Molecular Immunology database.

3.1.2. The Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database at Stanford (HIVRT & PrDB)

HIV protease is responsible for the cleavage of the polypeptide that codes for the structural proteins and enzymes of the virus. Reverse transcriptase is the enzyme responsible for converting viral RNA into DNA that can be incorporated into the human genome. These two proteins are the main targets for anti-retroviral (ARV) drugs aimed at preventing the human immunodeficiency virus from replicating within the human body. The significance of studying the HIV pol gene coding for reverse transcriptase and protease lies in the biological importance of these proteins and the severe effect of mutation on drug susceptibility (Kuiken et al, 2003). Of the 19 USA FDA approved ARVs seven are aimed at blocking reverse transcriptase, 7 target protease, and only one prevents HIV from fusing with and entering a human cell (http://www.thebody.com/atn/393/drugs.html).

The HIV reverse transcriptase and protease sequence database is an online database with the aim of connecting the nucleic acid sequence and additional data that can shed light on variation between the sequences of patients treated with anti-retrovirals (Shafer, Jung & Betts. 2000, Shafer et al. 2000, Kantor et al. 2001). The database contains a comprehensive collection of published reverse transcriptase and protease sequences both from the existing sequence databases, such as GenBank, and journal articles (Rhee et al. 2003, Shafer, Stevenson & Chan. 1999). Future improvements planned for the HIVRT & PrDB include the addition of gp41 sequences and data on resistance to fusion inhibitors.

Most database queries result in lists of information in tabular format. Alignments can be viewed in the traditional way or as a composite alignment (which can be obtained from http://hivdb.stanford.edu/cgi-bin/PRMutSummary.cgi). The alignments lack any graphical representations to enhance insight into the displayed information.

3.1.3. The HIV Resistance Response Database Initiative (RDI)

The aim of the RDI (http://www.hivrdi.org) is to standardize and improve the interpretation of genotypic information and to relate the genomic sequence changes in patients to the responses observed towards different HIV drug combinations. The genes coding for reverse transcriptase and protease are the foci of this project. The underlying database has been developed according to the conceptual schema (as used in data marts) to enhance performance when a query is submitted. HIV sequence and patient data are currently collected from the USA Military HIV Research Program, the BC Center for Excellence in HIV/AIDS, Vancouver, Canada, and the Italian HIV Cohort. Using the genotype, statistical and analytical methods are employed to accurately predict viral load response to therapy. The RDI project does not have a browsable presentation of the genetic data in the database.

3.1.4. The Africa Centre Molecular Virology and Bioinformatics Programme

The Africa Centre Molecular Virology and Bioinformatics Programme hosts a Website at http://www.bioafrica.net where a wide range of tools and resources for use in HIV research can be found, including the Genetic Data Environment (GDE) for HIV sequence analysis (De Oliveira et al, 2003). These tools and resources include subtype geographic maps, Bioinformatics software and proteomic resources. RNA genome maps of HIV-1, HIV-2 and SIV are available at http://www.vision.ime.usp.br/~tulio/ as a static PDF file. There is no interactive genome browser application available.

3.2. Overview of Available Genome Browsers

Although all of the browsers that will be discussed were originally designed and developed for eukaryotic genome analyses, it is worthwhile to consider them as the basis for an HIV browser. In the development of software packages for use in genomics and post-genomics, there are two routes to take: (1) develop software from the very beginning or (2) modify existing sofware to suit specific needs.

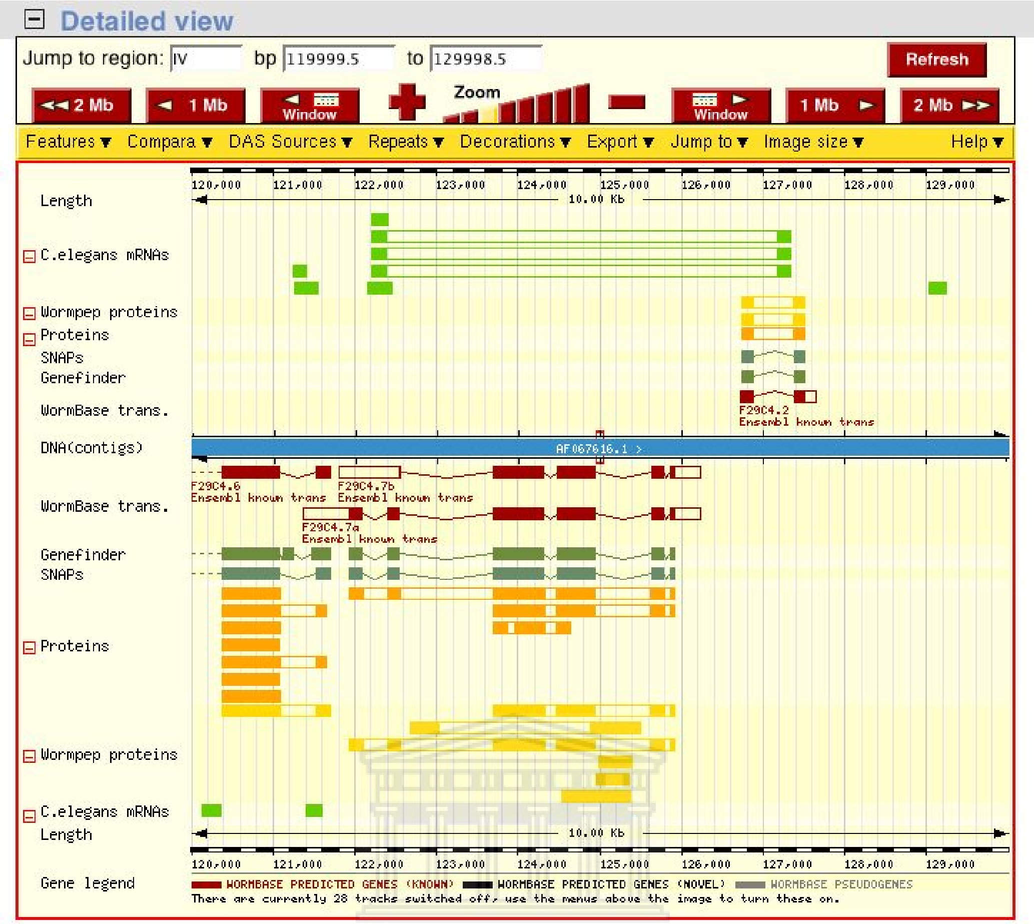
A brief overview of the various browsers (in alphabetic order) is given followed by an evaluation according to criteria defined in Chapter 2.

3.2.1. The Ensembl Genome Database Project

Ensembl (Birney et al. 2004, Hubbard et al. 2002, Kasprzyk et al. 2004) is the product of a collaborative effort between the European Bioinformatics Institute and the Sanger Centre. The goal of the Ensembl project can be subdivided into four categories: (i) to store and retrieve genome scale data, (ii) to serve as a Website for genome display, (iii) to provide an automatic annotation method for eukaryotic genomes, and (iv) to aid in the study of comparative genomics.

Ensembl is an Open Source project and relies only on Open Source software that is readily available and free of charge. The API (application programming interface) was originally written mainly in Perl and BioPerl but a JAVA API is now available and C extensions have been written. The Ensembl project makes use of relational MySQL databases. The database schema is very complex and contains a multitude of tables. The data for each available species is stored in a core database together with a number of other additional databases containing SNP, EST, haplotypes and other information. Information can be obtained by means of SQL queries to the MySQL server using the API or using the Web based GUI.

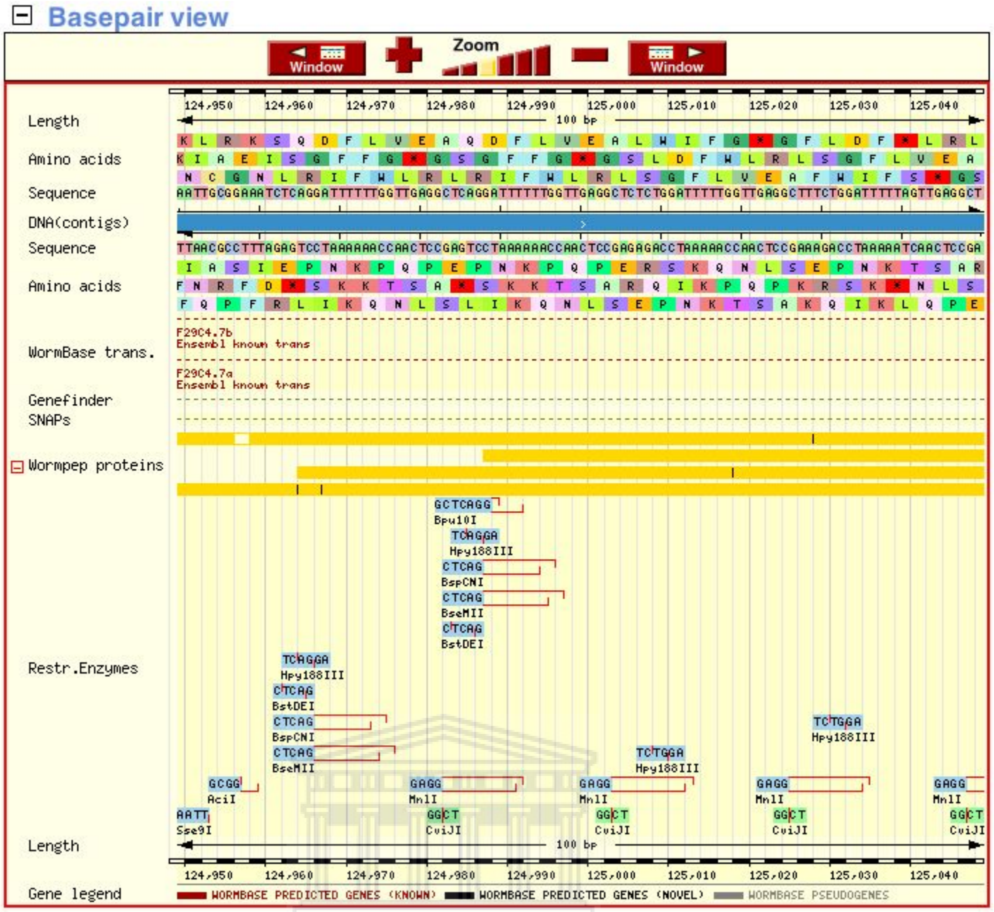
The Ensembl Genome Browser provides access to a number of different views. These views focus on specific areas of interest, for example ContigView allows the user to scroll along entire chromosomes to view available annotations while SyntenyView provides a view of blocks of conserved gene order between two species (Figure 3-5 and Figure 3-6). Most annotations are clickable glyphs that link to pages containing detailed information about the selected feature. The data available can be downloaded in a number of supported formats for use in external analyses. The data export feature is provided through ExportView.



Jump to region: IV bp 119999.5 to 129998.5

There are currently 28 tracks switched off, use the menus above the image to turn these on.

Figure 3-5 A screenshot of the ContigView of a region in the C. elegans' genome displaying a number of tracks and features of the browser such as scrolling, zooming and searching.



Basepair view
Window
Zoom
Window
Length
124,950 124,960 124,970 124,980 124,990 125,000 125,010 125,020 125,030 125,040
100 bp
Amino acids
K L R K S Q D F L V E A Q D F L V E A L W I F G * G F L D F * L R L
K I A E I S G F F G * G S G F F G * G S L D F W L R L S G F L V E A
N C G N L R I F W L R L R I F W L R L S G F L V E A F W I F S * G S
Sequence
AATTGCGGAAATCTCAGGATTTTTTGGTTGAGGCTCAGGATTTTTTGGTTGAGGCTCTCTGGATTTTTGGTTGAGGCTTTCTGGATTTTTAGTTGAGGCT
DNA(contigs)
Sequence
TTAACGCCTTTAGAGTCCTAAAAAACCAACTCCGAGTCCTAAAAAACCAACTCCGAGAGACCTAAAAACCAACTCCGAAAGACCTAAAAATCAACTCCGA
Amino acids
I A S I E P N K P Q P E P N K P Q P E R S K Q N L S E P N K T S A R
F N R F D * S K K T S A * S K K T S A R Q I K P Q P K R S K * N L S
F Q P F R L I K Q N L S L I K Q N L S E P N K T S A K Q I K L Q P E
WormBase trans.
F29C4.7b
Ensembl known trans
F29C4.7a
Ensembl known trans
Genefinder
SNAPs
Wormpep proteins
Restr.Enzymes
GCTCAGG
Bpu10I
TCAGGA
Hpy188III
CTCAG
BspCNI
CTCAG
BseMII
CTCAG
BstDEI
TCAGGA
Hpy188III
CTCAG
BstDEI
CTCAG
BspCNI
CTCAG
BseMII
TCTGGA
Hpy188III
TCTGGA
Hpy188III
GCGG
AciI
GAGG
MnlI
GAGG
MnlI
GAGG
MnlI
GAGG
MnlI
AATT
Sse9I
GGCT
CviJI
GGCT
CviJI
GGCT
CviJI
GGCT
CviJI
Length
100 bp
124,950 124,960 124,970 124,980 124,990 125,000 125,010 125,020 125,030 125,040
Gene legend
WORMBASE PREDICTED GENES (KNOWN)
WORMBASE PREDICTED GENES (NOVEL)
WORMBASE PSEUDOGENES

Figure 3-6 The sequence region under investigation can be viewed at DNA and amino acid level. A six-frame translation is provided. Restriction enzyme cut-sites can be seen at the bottom of the image.

Although Ensembl satisfies a wide range of the criteria that are important for the development of the HIV Genome Browser, the design was found to be too complex for our purposes. The type of data that will typically be stored in a molecular HIV database will not be able to fully justify such a heavy schema. In personal communications with X.M Fernández from Ensembl and M. Wilkinson (involved in the development of a Food and Mouth Disease Virus Genome Browser), it was determined that the Ensembl schema is too complex for the aim of the HIV Genome Browser.

3.2.2. The Generic Genome Browser (GBrowse)

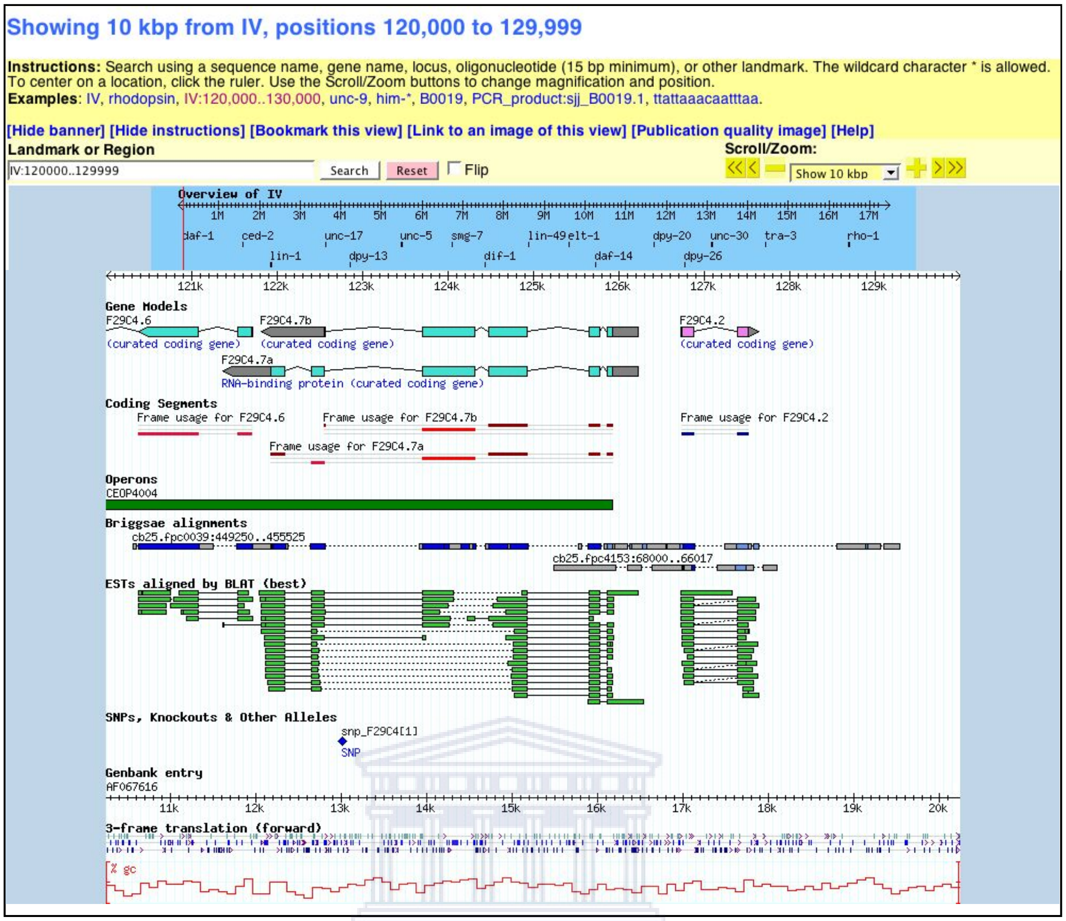
The Generic Genome Browser (Stein et al., 2002) was developed as a component of the Generic Model Organism Database (GMOD) whose goal is to develop component-based applications that can be used in separate model organism database systems. GBrowse is a visualization tool that provides a graphical representation of a given genome together with annotations and features relative to this genome. GBrowse was developed through the collaborative effort of members from different research groups. The collaboration ensured that the software would be portable between diverse model organism database systems. It is currently used to display genomes and annotations of Wormbase (http://www.wormbase.org), Flybase (http://www.flybase.org), Hapmap, *E.coli* and many other institutions. Figure 3-7 displays an example of the implementation of GBrowse using the genome data of *C. elegans*.

GBrowse is freely available and can be downloaded from http://www.sourceforge.net/project/showfiles.php?group_id=27707. It is relatively easy to adapt to suit a research group's specific needs. The source code can be downloaded and modified or extended and the help of the core developers and other users is easily accessible via the mailing list at gmod-gbrowse-request@lists.sourceforge.net. GBrowse is released under the Perl Artistic License. It is entirely built on top of freely accessible Open Source software and can use an Oracle, MySQL or PostgreSQL relational database for storage. The API was developed using Perl and BioPerl modules.

The database schema is simple and consists of seven tables accessed by a Perl API that links the CGI script generating the Web pages to the database. The tables can easily be loaded using a Perl script that comes bundled with the software. Data should be in a tab-delimited file according to the Gene Feature Format (GFF) described at http://www.sanger.ac.uk/Software/formats/ GFF/GFF_Spec.shtml.

The Web pages are highly customizable according to the developer's specifications and the type and number of tracks can vary between different installations. The availability of plugins ensures that GBrowse is very extendable and able to support new features as the need arises.

Displayed information can be downloaded in a number of formats to be used in external analyses. High quality images of the view currently displayed on the screen can be obtained by clicking on the link "*Publishable views*".



Showing 10 kbp from IV, positions 120,000 to 129,999

Instructions: Search using a sequence name, gene name, locus, oligonucleotide (15 bp minimum), or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.
Examples: IV, rhodopsin, IV:120,000..130,000, unc-9, him-*, B0019, PCR_product:sjj_B0019.1, ttattaaacaatttaa.

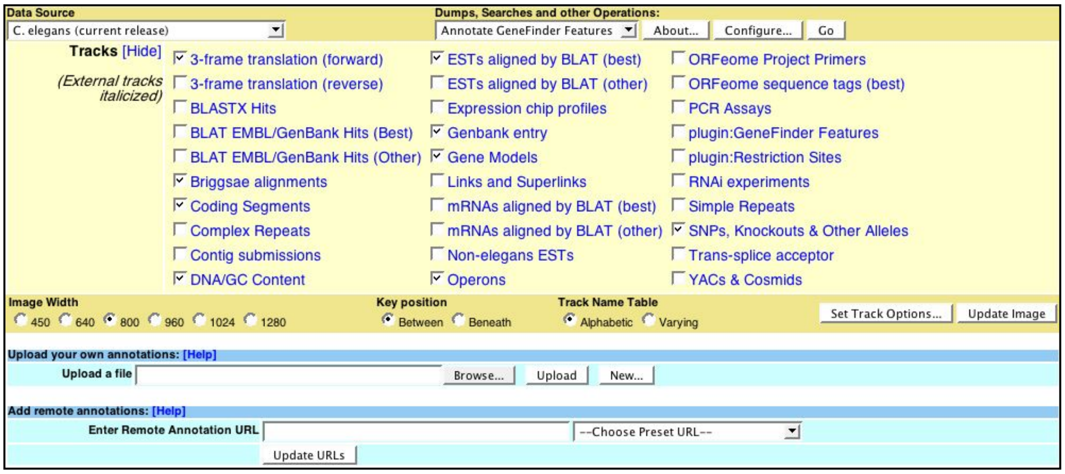
[Hide banner] [Hide instructions] [Bookmark this view] [Link to an image of this view] [Publication quality image] [Help]

Landmark or Region: IV:120000..129999 Search Reset □ Flip

Scroll/Zoom: Show 10 kbp

Figure 3-7 A screenshot of GBrowse as implemented by Wormbase. Scroll, zoom and search functionalities are available for the researcher to navigate to an area of interest. The display is divided into an overview showing the complete chromosome IV, and a detaile d panel where relevant annotation tracks are displayed. Different colours and glyphs are used to represent distinct features.

The user can upload his/her own annotations once the data is saved in the GFF format or a reduced three-column version thereof. The name of the file containing annotations with the recognized reference genome as its index can be entered in the space labelled "Upload a file" in Figure 3-8 and proprietory annotations will only be visible to the appropriate user. GBrowse also supports the DAS protocol.



Data Source

C. elegans (current release)

Dumps, Searches and other Operations:

Annotate GeneFinder Features | About... | Configure... | Go

Tracks [Hide]

(External tracks italicized)

☑ 3-frame translation (forward)
☐ 3-frame translation (reverse)
☐ BLASTX Hits
☐ BLAT EMBL/GenBank Hits (Best)
☐ BLAT EMBL/GenBank Hits (Other)
☑ Briggsae alignments
☑ Coding Segments
☐ Complex Repeats
☐ Contig submissions
☑ DNA/GC Content

☑ ESTs aligned by BLAT (best)
☐ ESTs aligned by BLAT (other)
☐ Expresion chip profiles
☑ Genbank entry
☑ Gene Models
☐ Links and Superlinks
☐ mRNAs aligned by BLAT (best)
☐ mRNAs aligned by BLAT (other)
☐ Non-elegans ESTs
☑ Operons

☐ ORFeome Project Primers
☐ ORFeome sequence tags (best)
☐ PCR Assays
☐ plugin:GeneFinder Features
☐ plugin:Restriction Sites
☐ RNAi experiments
☐ Simple Repeats
☑ SNPs, Knockouts & Other Alleles
☐ Trans-splice acceptor
☐ YACs & Cosmids

Image Width: ○ 450 ○ 640 ● 800 ○ 960 ○ 1024 ○ 1280

Key position: ● Between ○ Beneath

Track Name Table: ● Alphabetic ○ Varying

Set Track Options... | Update Image

Upload your own annotations: [Help]

Upload a file | Browse... | Upload | New...

Add remote annotations: [Help]

Enter Remote Annotation URL | --Choose Preset URL--

Update URLs

Figure 3-8 The variety of tracks that can be selected in order to customize the display of the WormBase browser. Space is provided at the bottom of the screen to upload files or supply URLs containing private annotations.

3.2.3. NCBI Map Viewer

The Map Viewer available at the NCBI Website (http://www.ncbi.nlm.gov/mapview) provides graphical displays of certain genome assemblies available in Entrez genomes (Wheeler et al. 2003). It currently provides integrated views of chromosome maps for 17 organisms.

Map Viewer is not completely browser independent. It can be viewed best in recent versions of Netscape and Internet Explorer on PCs, and Internet Explorer on Macs.

Regions of interest can be located through a text search or by specifying the coordinates of the region in the genome under investigation. Chromosome maps are displayed and can be zoomed in on to reveal more detail (Figure 3-9). At the highest resolution the DNA sequence can be viewed together with annotations including CDS regions, RNA coding regions, and predicted genes.

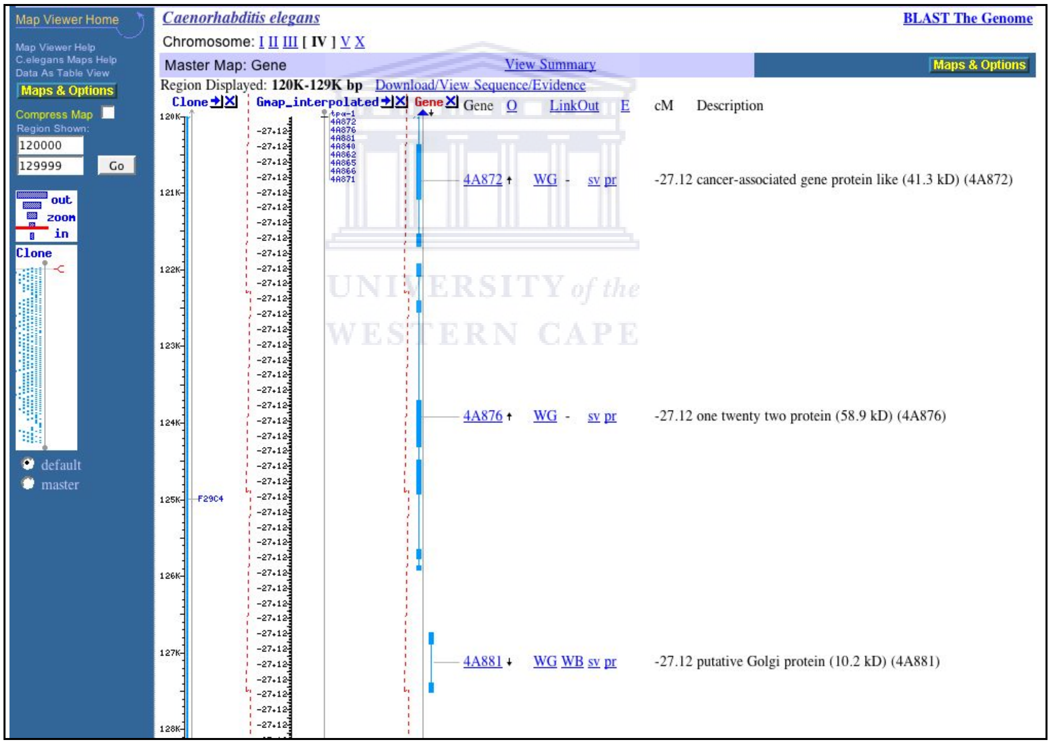
When viewing the sequence, the user can navigate through the chromosome by means of a scroll button. It is also possible to zoom in or out depending on the requirements of the user. The option to view the corresponding region on the minus strand is available. If the user clicks on a gene or other region of interest the view is changed to include only the relevant gene.

The view can be customized by the user through a link named "Maps & Options". The user can choose which maps to display and in which order. Map Viewer supports comparative viewing of different genomes. Users can select which other genomes to display relative to the original genome of choice.

The graphical view cannot be saved or downloaded. The data displayed in the map is available as a tab-delimited file, called the Table View, for download. The resulting file contains the start and stop positions of the annotations in the area of interest, the gene names, strandedness (+ or -), links to more information about the specific genes, the graphical display in Map Viewer (Figure 3-10), from genes to RefSeq proteins, and a description of gene products. The sequence region covered by the current display can also be downloaded as a FASTA file or in GenBank format.

Although Map Viewer only provides access to a limited number of complete genomes, over 800 complete genomes are available through Entrez. Entrez genomes provides a graphical overview of these genomes together with some annotations through simple browsing features. Only the sequence view of Map Viewer (as described above) can be obtained for these genomes.

Because Map Viewer forms an integrated part of a larger non-portable system and cannot be redistributed it was found not to be useful as a basis for the HIV genome browser.



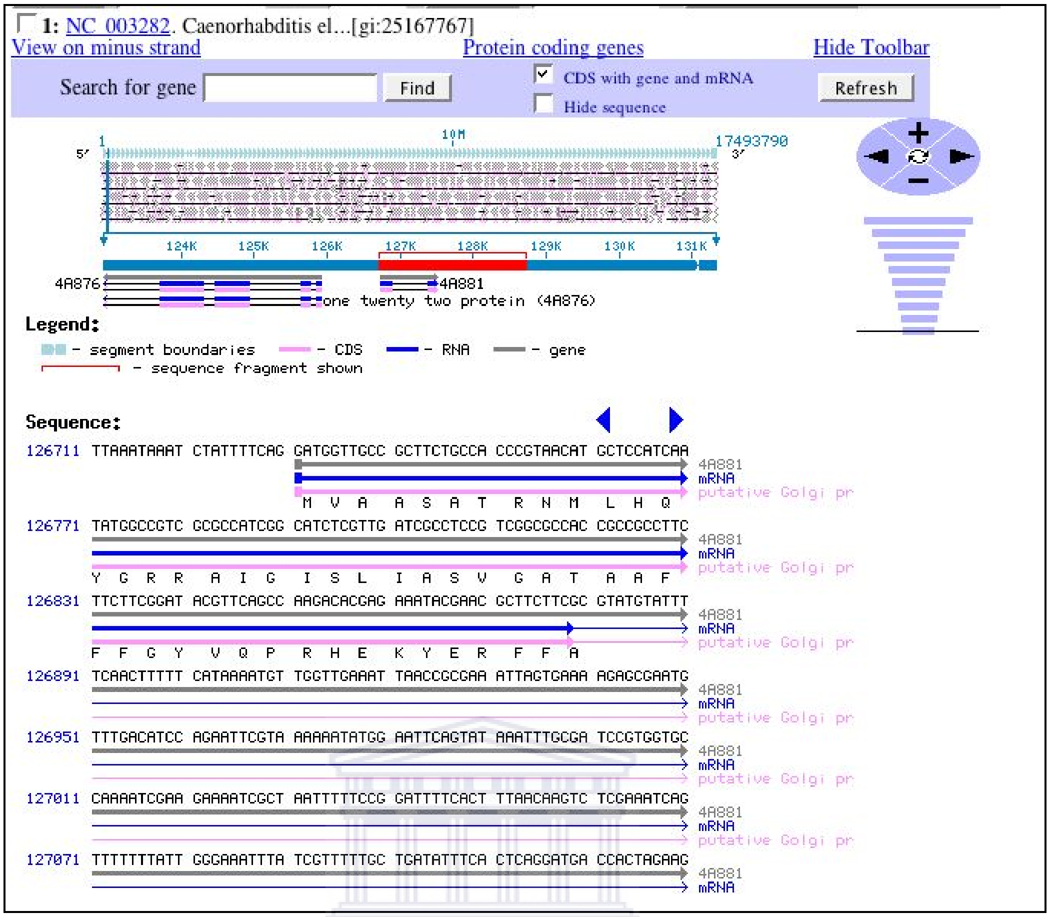
Master Map: Gene

Region Displayed: 120K-129K bp

Region Shown: 120000 129999 Go

Gene	O	LinkOut	E	cM	Description
4A872 ↑	WG	-	sv pr		-27.12 cancer-associated gene protein like (41.3 kD) (4A872)
4A876 ↑	WG	-	sv pr		-27.12 one twenty two protein (58.9 kD) (4A876)
4A881 ↓	WG	WB	sv pr		-27.12 putative Golgi protein (10.2 kD) (4A881)

Figure 3-9 A screenshot of chromosome IV bases 120000 to 129999 of the C. elegans' genome in Map Viewer.



Legend:

▨ - segment boundaries, ▬ - CDS, ▬ - RNA, ▬ - gene

⌐¬ - sequence fragment shown

Sequence:

```
126711 TTAAATAAAT CTATTTTCAG GATGGTTGCC GCTTCTGCCA CCCGTAACAT GCTCCATCAA
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
                             M   V   A   A   S   A   T   R   N   M   L   H   Q
126771 TATGGCCGTC GCGCCATCGG CATCTCGTTG ATCGCCTCCG TCGGCGCCAC CGCCGCCTTC
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
       Y   G   R   R   A   I   G   I   S   L   I   A   S   V   G   A   T   A   A   F
126831 TTCTTCGGAT ACGTTCAGCC AAGACACGAG AAATACGAAC GCTTCTTCGC GTATGTATTT
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
       F   F   G   Y   V   Q   P   R   H   E   K   Y   E   R   F   F   A
126891 TCAACTTTTT CATAAAATGT TGGTTGAAAT TAACCGCGAA ATTAGTGAAA AGAGCGAATG
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
126951 TTTGACATCC AGAATTCGTA AAAAATATGG AATTCAGTAT AAATTTGCGA TCCGTGGTGC
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
127011 CAAAATCGAA GAAAATCGCT AATTTTTCCG GATTTTCACT TTAACAAGTC TCGAAATCAG
                                                                    4A881
                                                                    mRNA
                                                                    putative Golgi pr
127071 TTTTTTTATT GGGAAATTTA TCGTTTTTGC TGATATTTCA CTCAGGATGA CCACTAGAAG
                                                                    4A881
                                                                    mRNA
```

Figure 3-10 A link is provided in Map Viewer to a horizontal display of the region of interest together with gene, CDS and RNA annotations. The scroll or zoom buttons or the search box can be used to navigate through the chromosome. The DNA sequence and amino acid translation is provided.

3.2.4. The UCSC Genome Browser

The goal as described by Kent *et al.* (2002) can be summarized as the provision of an integrated graphical representation of the human genome and theoretically predicted and experimentally confirmed annotations thereof, in order to be able to relate sequence information to biological knowledge. It also facilitates comparative genomics since it now, in addition to the human genome sequence, provides the reference sequence for *C. elegans*, and working drafts for the mouse, rat, Fugu, and *C. briggsae* genomes. Recently it was announced that the SARS coronavirus TOR2 draft assembly has been made available at the UCSC Genome Browser site. This is significant, since all the other genomes that are currently available in genome browsers, are eukaryotic.

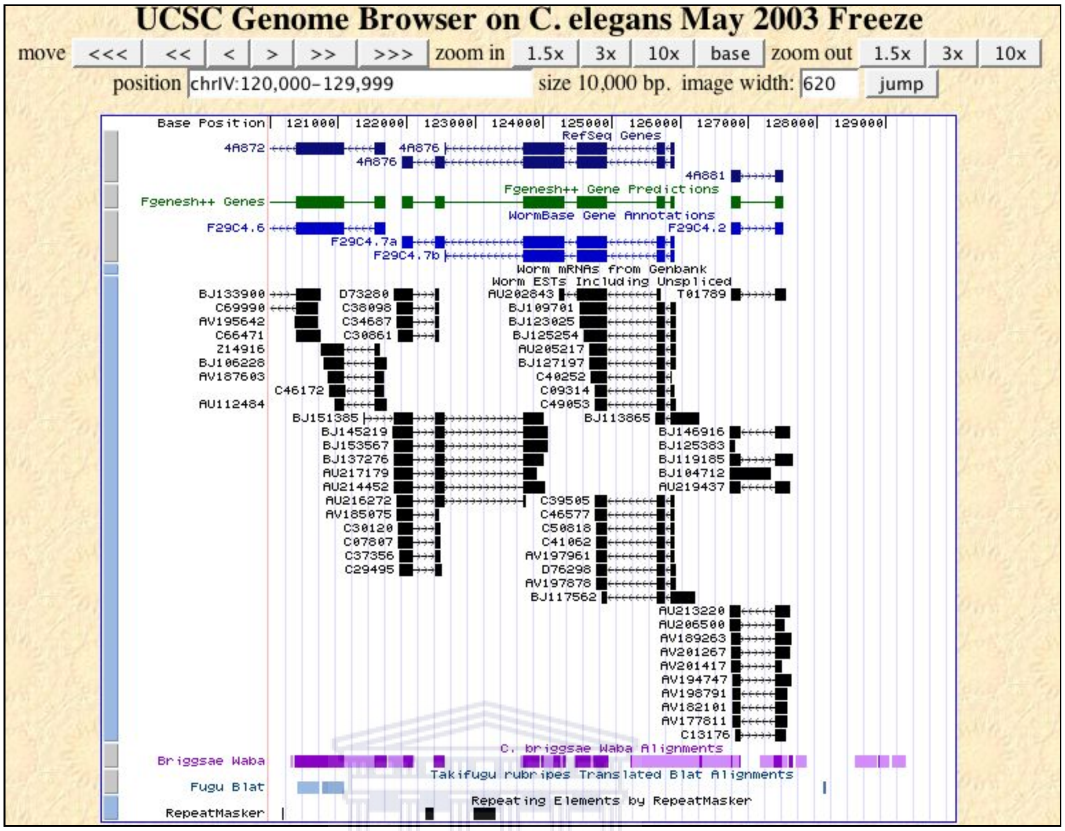
The Browser has an HTML/CGI front end. The CGI source code is written in C. The database (Kent *et al.*, 2002) upon which the browser is built is a relational MySQL database. Tables are divided into two groups: those that provide positional information and those that contain non-positional information such as DNA sequence, author, source, etc. (which is not displayed). The latter provides information that is useful when more detail about a specific genomic region is required. The entire database can be downloaded as tab-delimited files from http://genome.ucsc.edu and is updated weekly. It is also possible to download specified sub-sets of the data through the use of a "Table Browser" (http://genome.ucsc.edu/goldenPath/hgText.html).

The middle layer between the browser and the database is a program written by Jim Kent called **autoSql**. **AutoSql** is a hybrid between SQL and the C programming language and is used to populate the database and plays a role in memory management. Because a hybrid of C and SQL is used to program the middleware and user interface layers, it is more difficult to adapt this system to work with other software/projects where the main programming languages include Perl, BioPerl, Python and Java.

Displayed graphics can be downloaded in either PDF or Postscript format. To obtain sequence and feature information the researcher can choose to download the data either as a tab-delimited file or as a FASTA-format text file. Other options for sequence manipulation include reverse complementation, switching to upper/lower case and applying repeat masking. It is also possible to export the sequence in the form as an HTML file. Figure 3-11 shows the availability of scroll and zoom functions and a display of some features present in the *C. elegans* genome in the UCSC browser. Figure 3-12 demonstrates the various tracks that can be viewed by a researcher interested in specific features of the *C. elegans'* genome.

Researchers are able to upload their own annotations through the Genome Browser's custom annotation track feature. As seen in GBrowse, these annotations are not permanent or public and will only be available for a limited amount of time on the machine where the data resides. Individuals can also publicize an URL in order to make their annotations available to other researchers or collaborators. UCSC supports the DAS protocol.

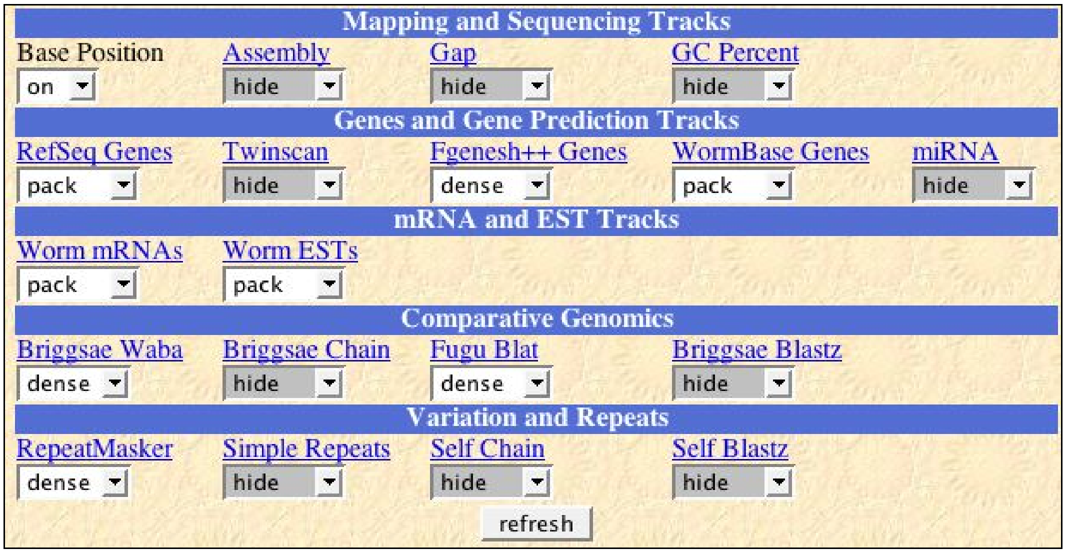
The database and browser at UCSC could be extended to contain the genomic sequences of more species, and to view a wider range of annotation tracks than is currently available.



UCSC Genome Browser on C. elegans May 2003 Freeze

position chrIV:120,000-129,999 size 10,000 bp. image width: 620

Figure 3-11 A view of a region in the C. elegans' genome as displayed by the UCSC Genome Browser. The scroll, zoom and search functions are found at the top of the image.



Mapping and Sequencing Tracks

Base Position	Assembly	Gap	GC Percent	
on	hide	hide	hide	

Genes and Gene Prediction Tracks

RefSeq Genes	Twinscan	Fgenesh++ Genes	WormBase Genes	miRNA
pack	hide	dense	pack	hide

mRNA and EST Tracks

Worm mRNAs	Worm ESTs
pack	pack

Comparative Genomics

Briggsae Waba	Briggsae Chain	Fugu Blat	Briggsae Blastz
dense	hide	dense	hide

Variation and Repeats

RepeatMasker	Simple Repeats	Self Chain	Self Blastz
dense	hide	hide	hide

Figure 3-12 In the UCSC Genome Browser tracks can be turned on or off. The display can also be dense or packed, in which case details of seperate features of a track are not displayed.

3.2.5. Other Approaches to Genome Visualization

Other approaches have been taken to visualization of genomes and these are discussed briefly below. One approach is a graphical display of genome comparisons. Through such a display homology can be detected in different genomes and rearrangements and inversions can be identified. BugView (Leader, 2004) is a browser based on this comparison approach. It is a Java application for visualizing either comparisons of two genomes or a single genome. It was developed for visualizing bacterial genomes, but can also be applied to eukaryotic genomes (maximum size 30Mb). BugView displays the genomes vertically with a scale to indicate relative position and highlights linking homologues between the genomes being compared.

Phylo-VISTA (Shah et al, 2004) is another Java application for the comparison of different genomes. It differs from BugView in that multiple DNA sequence alignments rather than pairwise only genome comparisons can be viewed. Phylogenetic trees are also drawn to describe phylogenetic relationships between different sequences.

In addition to viewing similarity results between sequences, viewing polymorphisms is a very important approach that has been addressed by viewGene (Kashuk et al, 2001). viewGene is a standalone Java application that runs on Windows, Solaris, Linux and Macintosh machines. The display is divided into three sub-windows: the *Features* window includes all annotations as typically found in a GenBank file (other annotations obtained from external analysis packages can also be included); the *Matches* window includes data obtained via electronic means such as FASTA or BLAST searches, and the *Fragments* sub-window displays sequences obtained from the local laboratory that match the region under investigation. Polymorphisms can be viewed in the *Matches* and *Fragments* sub-windows. The user can select which base changes are shown, for example showing only G→A changes occurring in exons. At the time of publication viewGene was optimized to display up to one megabase of sequence. The figures that are generated can be saved as GIF or JPG images.

Another approach to genome visualization is that of an annotation viewer and/or editor. Examples annotation viewer/editor systems include Apollo, Genquire, and Artemis.

Apollo (Lewis et al, 2002) was developed to improve computational annotations through the manual intervention of expert curators. Apollo allows researchers to access genome annotations in a graphical view and edit them where needed. Both the forward and reverse strands of a genome and all relevant annotations to both orientations can be seen. Zooming and scrolling are provided to navigate through the genome. Apollo is written in Java.

Artemis (Rutherford et al, 2000) is another Java-based sequence annotation tool developed for annotation and editing of microbial and lower eukaryotic genomes. Results from external analysis programs can be overlaid on the genome under investigation and annotations can be modified according to evidence from such analyses. As in

Apollo, both orientations (forward and reverse) can be viewed with all annotations and a 6-frame translation of the codons. It is possible to zoom in and out to show or hide detail, ranging from only stop/start codons to the complete nucleic acid or amino acid sequence.

Genquire (Wilkinson et al, 2002) is a hybrid between a genome annotation tool and a genome browser. It allows querying of a genome available from the underlying database or supporting flat file as with other genome browsers, but it also supports editing of annotations. Genquire is a standalone application implemented in Perl/Tk using BioPerl methods. It offers three different views: the Genome Map, which displays chromosomes in a vertical fashion; the horizontally or vertically displayed Sequence Canvas showing all sequence features; and the Nucleotide Sequence Viewer that displays the nucleotide sequence with regions of interest identified.

3.3. Evaluation According to Defined Criteria

To perform a thorough evaluation of the various browsers according to criteria set in Chapter 2, the source code was examined (where available), data schemas were studied, Web sites were rated, and installation instructions were reviewed.

Table 3-1 and Table 3-2 summarizes the evaluation of the browsers with regards to the criteria of Chapter 2. Requirements were rated as “yes/no” or “√/x”. A discussion of results displayed in Table 3-1 and Table 3-2 follows.

Developer Criteria	**Ensembl**	**GBrowse**	**Map Viewer**	**UCSC**
Web-based	√	√	√	√
Sufficient documentation	√	√	N/A	x
Easy installation	x	√	N/A	x
Modular	√	√	N/A	√
Support from original developers	√	√	x	√
Portable	x	√	N/A	x
Extensible	√	√	N/A	x
Easy to maintain	x	√	N/A	x
Uses standardized protocols/language	x	√	N/A	x
Open-source	√	√	x	√

Table 3-1 A comparison of four existing genome browsers according to criteria defined by the developer.

User Criteria	Ensembl	GBrowse	Map Viewer	UCSC
User-friendly	√	√	√	√
Intuitive user interface	x	√	√	√
Platform independent	√	√	√	√
Access to current data	√	√	√	√
Low cost to obtain and maintain	√	√	√	√
Secure	√	√	N/A	N/A
Highly responsive	x	√	√	√
Interactive	√	√	√	√
DAS support	√	√	x	√
Download display for further analyses	√	√	√	√
Publishable view support	√	√	x	√
Good organization of data display on screen	x	√	x	√
Accurate display of data	√	√	√	√
Semantic zooming supported	x	√	x	√
Support wide range of queries	√	√	√	√
Customizable display	√	√	√	√
View private data	√	√	x	x
Provide detailed additional information (or links to such information) on regions of interest	√	√	√	√
Display multiple sequence entries simultaneously	√	x	√	√
Display information relevant to a set of sequences	√	x	x	√

Table 3-2 An evaluation (from the user's perspective) of four currently available genome browsers according to criteria set in Chapter 2.

3.3.1. Developer Criteria

In Chapter 2 the reasons for choosing a Web-based system above a stand alone application was already elucidated. All four of the browsers evaluated in this study were Web-based applications.

Documentation was regarded as being sufficient (in the context of this study) if clear instructions for downloading, installation and implementation of a system was available. Additional information regarding dependencies of the software under investigation would benefit the developer and was rated as a very important component of "sufficient documentation". In this category only Ensembl and GBrowse were satisfactory.

Access to help from the developer community of a specific system supported ease of installation and implementation. The GBrowse developers are extremely helpful and usually reply to emails sent to the mailing lists within 24 to 48 hours. Ensembl and UCSC genome browser developers were also highly responsive.

Contrary to Ensembl, GBrowse and the UCSC genome browsers, NCBI's Map Viewer cannot be installed locally. Ensembl and the UCSC browser are mainly distributed to serve as mirror sites of databases distributed by the developers. GBrowse, however, can be installed locally and used to display proprietary (local) data. A database with the general GBrowse schema has to be created but can be populated by data as chosen by the user. Sample data are , however, distributed with the GBrowse package to test the installation.

Installation of the Ensembl and UCSC databases and viewers is very complex and time consuming. Ensembl uses some non-standard BioPerl modules and an old version of Apache. Database downloads are substantial. The GBrowse distribution is about 1.35 MB and installation can be done within an hour. The user does not have to download large amounts of data to test the browser and can immediately start to implement his/her genome of interest. In some cases to generate the GFF files with which GBrowse tables are populated can be extremely laborious but scripts facilitating conversion of standard formats to GFF format is becoming available – an example includes *bp_genbank2gff.pl* which converts GenBank files to GFF.

Modularity, extensibility and portability are interconnected. GBrowse, Ensembl, and UCSC's genome browser are modular in that they all consist of a seperate database with a program on top of the database that renders the Web based front end. GBrowse is much more extensible and portable. It is also a much smaller package to download and can be installed on almost any PC without taking up too much diskspace or CPU. Functionality available through GBrowse can be extended easily through development of plugin perl scripts – that perform specific tasks. GBrowse functionality can also be extended through communication with the developers. Functions that could be useful to the wider GBrowse community can easily be incorporated into the main distribution with future releases.

Because of the size of Ensembl and the UCSC genome databases, the process of updating information available from these systems are very laborious (R. Brauning, personal communication). The level of ease of updating information available through GBrowse will depend on the data source (i.e. another database) that feeds the GBrowse database. In the proposed system for the HIV Genome Browser, the GBrowse database will be updated automatically from a database containing all relevant information via a SQL/perl script. Updating the distribution of GBrowse is very easy once it has been installed properly.

The APIs for both Ensembl and GBrowse is written in Perl and BioPerl. MySQL databases are used in the case of Ensembl and UCSC. Through the use of the correct database adaptor GBrowse can support MySQL, PostgreSQL, or Oracle databases. The UCSC browser front end is coded in C. The middle layer of the UCSC browser is written in autoSql – a hybrid between C and SQL written by Jim Kent. Although Ensembl implements a variety of BioPerl modules, these are non-standard and have to be obtained from the Sanger Centre's Web site and installed seperately to the normal distribution of BioPerl. GBrowse generally makes more use of commonly accepted Bioinformatics tools and protocols than the UCSC browser and Ensembl.

The source code for Ensembl, GBrowse and the UCSC Genome Browser can be downloaded from the respective URLs provided in section 3.2. They are free for academic use with specific restrictions described in their licenses.

3.3.2. User Criteria

One of the most important aspects of introducing a new software tool to the user community is user-friendliness. An application that requires hours of training and practice to use in a meaningful way is not welcomed into the research community. Users want applications that will add value to their results obtained through laborious efforts and dedication, without spending valuable hours trying to understand the user interface and available functions. All of the evaluated browsers qualified as being user-friendly. The Ensembl user interface, however, is heavily populated with data and could be overwhelming for new users. In most of the browsers it is possible to only display tracks of interest to the researcher - thus reducing screen clutter.

Various Web browsers, including Mozilla, Internet Explorer and Netscape, were successfully used with all of the genome browsers under investigation.

Access to current data is the responsibility of the genome browser database administrator (excluding Map Viewer). All of the browsers have the potential to display the latest available data for the genome of interest. It is published on a regular basis on the relevant Web sites for Ensembl and UCSC. Displaying current data in GBrowse is discussed in 3.3.1.

Security of data displayed in a browser is a very important point for researchers. Although problems were previously experienced with security in GBrowse, it has now been solved (L. Stein, personal communication). Users can upload their data into GBrowse either by supplying a file or URL containing relevant information, or by populating a GBrowse database, without the fear that other people might intercept their unpublished work.

Responsiveness of software applications should be high. It can be difficult to satisfy this criterion due to network constraints in South Africa. The complexity of the browser user interface can decrease responsiveness noticeably as in the case of Ensembl. The large number of graphics embedded in a typical Ensembl page can cause long page loads. The same problem was not experienced with any of the other browsers.

Users want to be able to ask a wide variety of questions from the database supporting a browser. Communicating with the database through a variety of entry points – keywords, coordinates, and short sequences – makes the interaction satisfactory for the researcher. All of the browsers seemed to be very interactive and supported a wide range of questions that could possibly be asked by the user. Display of data across the screen was generally accomplished better by GBrowse and the UCSC browsers.

DAS support is provided by Ensembl, the UCSC genome browser and GBrowse. Users can view data sourced from a variety of available DAS servers on either of these browsers.

Displaying data available from a database underlying a genome browser will satisfy only some of the needs set by a researcher. Once certain qualities of the area of interest have been established through a graphical representation, the user will want to download sequence data spanning the region of interest to use in further analyses or studies. All of the browsers support the download of sequence data for the displayed region. Publishable views of the displayed region are also supported by all browsers but NCBI's Map Viewer.

A user can view private data in the Ensembl and GBrowse browser without having to load data into a database and without making it accessible to other users.

HIV researchers require display of a number of sequences simultaneously to estimate the level of diversity or to identify highly conserved regions. Ensembl and the UCSC browser fulfill this criterion, but the current distribution of GBrowse does not support the functionality. GBrowse can also not display data relevant to sets of sequences (e.g. positive selection or entropy).

All browsers under investigation supply links to where more information can be found on a region or feature of interest by clicking on such a feature.

3.3.3. Choice of a Basis for the HIV Genome Browser: GBrowse

GBrowse was chosen as a basis for development of the HIV Genome Browser Prototype, based on support of required biological functionality and the previously defined criteria for the browser. It performed better than any of the other systems according to results of the evaluation.

Although Gbrowse does not currently support any HIV-specific requirements, it allows the addition of plugin modules to extend it's usability. Plugin modules can be maintained seperately from the main program's code. Plugin modules can also be distributed with the main distribution to meet the needs of researchers working on other virusses.

Through personal communications with M. Wilkinson it was established that Gbrowse was also used to display Food and Mouth Disease Virus data while Ensembl was considered as being an over-kill for this purpose. Further justification for using GBrowse and the implementation of the HIV Genome Browser is discussed in Chapter 4.

Chapter 4 The HIV Genome Browser Prototype

4.1. A Case for Using GBrowse as an HIV Genome Browser

The Generic Genome Browser is a relatively simple component-based Web application that fulfills a great number of the requirements of the HIV Genome Browser. In Chapter 3 an evaluation of GBrowse according to predefined criteria was presented. According to these criteria GBrowse performed the best of the four candidate genome browsers that were assessed.

The application is open-source freeware and is easy to obtain and install. The software dependencies can all be obtained free-of-charge and are installed just as easily. The developer and user communities provide support and replies to questions or comments to the mailing list (gmod-gbrowse-request@lists.sourceforge.net) can generally be expected within 48 hours.

Creating a new database and populating it with data requires some knowledge of MySQL, the construction of a GFF file (either manually or via a provided Perl script), and running of a Perl script to load data into the database. Examples of both GFF files and configuration files are provided and can be copied to ensure the correct format - especially in the case of the configuration files.

Gbrowse allows the display of a very wide range of features, and new glyphs to display features more accurately are being developed continuously. It is relatively easy to upload new reference sequences and annotations into the database once it has been created.

The functionality of the browser can be extended either by writing plugins and placing them in the gbrowse.conf directory or by requesting the addition of new features from the developers. If the requested functionality is deemed important and beneficial to the wider user community, it will be added in future releases of the software.

In the case of an HIV genome browser an example of a plugin that could be written is the display of accumulated epitopes per patient over time. The same could be done for mutation accumulation at different time points. Visual displays of epitope and mutation accumulation may support identification of potential vaccine candidates. Because of time limitations these plugins have not been written, but some thought has gone into scripts that could display the above-mentioned data.

Although GBrowse seems to be a good starting point for the development of an HIV genome browser, it does have limitations. Because it needs a reference genome upon which to layer annotations, and because it can only display one reference genome at a time, it is difficult to display sequences from, for example, different patients simultaneously. Sequences from different patients will each exist as independent reference coordinate systems since they have unique identifiers. It is, however, possible to have all sequences in a common reference system (e.g. through alignment to a reference such as

HXB2), but in this case the browser will display all information available for the selected reference system per activated track. It may be less informative to view all available data thus to make use of the above-mentioned solution the developer should find a way for the user to limit queries to display only data of interest.

Displaying information pertaining to sets of sequences is problematic, since GBrowse can't display sets of sequences unless they are aligned to the reference genome. The types of information included are: positive selection graphs, entropy graphs and graphs portraying the number of changes in amino acid sequence for a multiple alignment from the reference sequence. Positive selection and entropy graphs are not currently displayed in the browser but will be very valuable once a method to display it in a biologically meaningful way has been established.

All of these obstacles can be overcome since GBrowse is open-source and the CGI script that generates the Web pages can be customized to suit the needs of the HIV Genome Browser.

4.2. General Structure and Features of GBrowse

The different components of GBrowse will be discussed with an explanation of the local HIV Genome Browser implementation.

4.2.1. The GBrowse Schema and Its Different Components

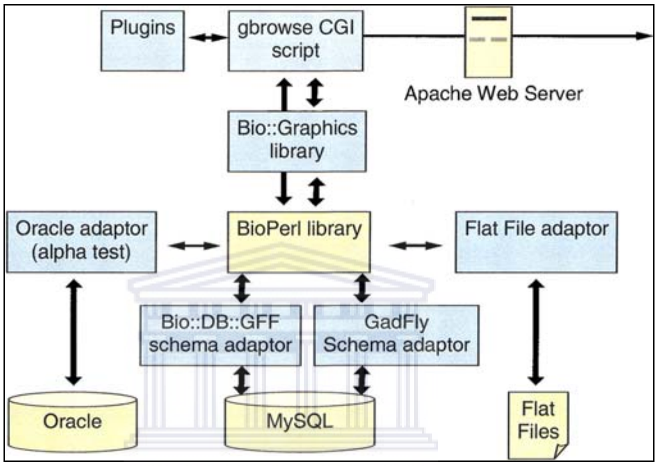


Figure 4-1 A representation of the Generic Genome Browser's schema (Stein et al, 2002).

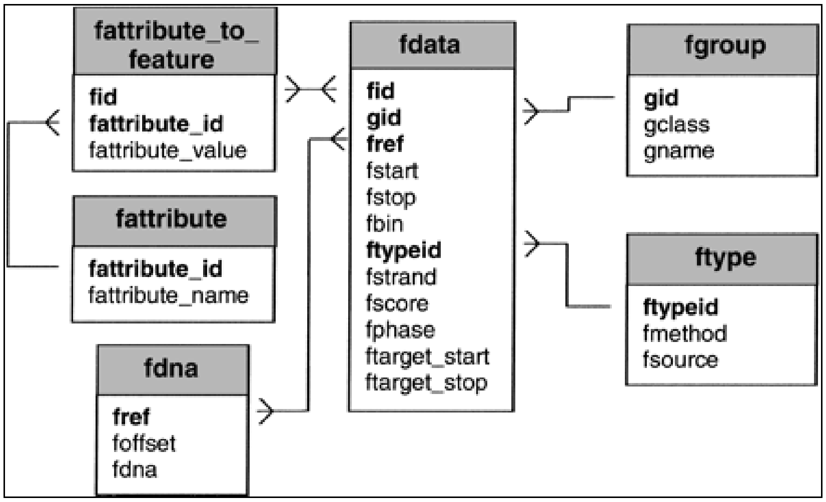


Figure 4-2 The seven tables that make up the Bio::DB::GFF schema (obtained from Stein et al, 2002).

Figure 4-1 shows a diagrammatic representation of the GBrowse schema. Data can be obtained from MySQL, PostgreSQL or Oracle databases or flat files. The MySQL database schema will be discussed in more detail since the HIV Genome Browser was built on a MySQL database. The database consists of seven tables (Figure 4-2). Perl scripts to load the database is available with the BioPerl installation. *bp_bulk_load_gff.pl* creates the tables and loads GFF annotations and FASTA sequence files in the appropriate tables. *bp_load_gff.pl* can be used to load additional data files into the database.

The BioPerl library is used to separate the database from the CGI script that creates the Web pages.

The GBrowse CGI script reads a stylesheet and configuration file used to generate the Web pages. The stylesheet contains information about the colour scheme and font sizes.

The configuration file is divided into two sections: the "general" section and the "tracks" section. In the general section the data source (either a database or flat file), and the database adaptor (Bio::DB::GFF) is specified. The general section contains compulsory and optional information. Specifying the reference class is essential. The reference class refers to the third column of the GFF file where the reference sequence is specified. A reference sequence must be specified in all GFF files and must match with the one set in the configuration file. Annotations are layered on top of a reference genome sequence according to their start and stop positions within the reference. Alternative "automatic classes" that will be searched if an unqualified identifier is given in the search box, for example "gag" instead of "gene:gag", can be specified. Automatic classes generally refer to other columns or information contained in the GFF file.

Plugins are Perl scripts that have been written by third parties to perform custom functions. These scripts reside in the plugin directory in the gbrowse configuration directoryand are specified in the configuration file. The path to the stylesheet, buttons and temporary images to generate the Web pages are also specified in the configuration file.

The second section of the configuration file contains the definitions of different tracks that will be displayed in the browser. Each track is defined by specifying the feature that it represents. The glyph is chosen from a defined selection available through the BioPerl Bio::Graphics module. Colour, labeling, and track name can also be specified. Depending on the glyph other parameters can be defined (for example the xyplot-glyph can have an axis, minimum and maximum values and either a line graph or a histogram-format graph). The URL associated with a feature can be set in the tracks section of the configuration file.

4.2.2. Implementation of the HIV Genome Browser

4.2.2.1. Software and Database Installation

GBrowse version 1.61 (released 2004-03-19) was obtained from http://prdownloads.sourceforge.net/gmod/Generic-Genome-Browser-1.61.tar.gz?download and installed on a dual Intel Xeon Server with 4 GB RAM. The operating system is FreeBSD 4.9. Additional required software packages which were installed are BioPerl 1.4, MySQL 4.0.14, Perl 5.8.2, Apache 1.3.

The HIV Genome Browser currently uses a MySQL database. After creating a MySQL database called HIV1, the GFF files (described in section 4.2.2.2.) were loaded using the *bp_bulk_load_gff.pl* and *bp_load_gff.pl* scripts provided with the GBrowse software distribution. The FASTA files containing viral sequences were loaded using the *bp_load_gff.pl* script. Complete HIV genome sequences were obtained from Entrez in GenBank format and loaded into the database using the Perl script *bp_genbank2gff.pl.* Figure 4-3 illustrates the commands used to create a MySQL database and to populate the newly created database.

```
(a)  lagbaja:~% mysql -uroot

       Welcome to the MySQL monitor.  Commands end with ; or \g.
       Your MySQL connection id is 154715 to server version: 4.0.14-log

       Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

       mysql> CREATE DATABASE HIV1;
       Query OK, 1 row affected (0.07 sec)

       mysql> GRANT ALL PRIVILEGES ON HIV1.* TO anelda@localhost;
       Query OK, 0 rows affected (0.05 sec)

       mysql> GRANT SELECT ON HIV1.* TO www@localhost;
       Query OK, 0 rows affected (0.00 sec)
```

```
b)   lagbaja:~% bp_bulk_load_gff.pl -user root -d HIV1
                DATABASE/HXB2_ANDallEpis.gff
     This operation will delete all existing data in database HIV1.  Continue? y
     Preparing embedded sequence....
     done....
     Loading feature data.  You may see duplicate key warnings here...
     done...
     2303 features successfully loaded

     lagbaja:~% bp_load_gff.pl -user root -d HIV1 DATABASE/UCT_Genbank2.gff
                DATABASE/UCT_Genbank2.fasta
     DATABASE/UCT_Genbank2.gff: loading...
     DATABASE/UCT_Genbank2.gff: 94 records loaded
     Loading fasta file DATABASE/UCT_Genbank2.fasta
     DATABASE/UCT_Genbank2.fasta: 15 records loaded
     lagbaja:~% bp_load_gff.pl -user root -d HIV1 DATABASE/mutdens.gff
     DATABASE/mutdens.gff: loading...
     DATABASE/mutdens.gff: 7 records loaded

     lagbaja:~% bp_load_gff.pl -user root -d HIV1 DATABASE/abepitopedens.gff
     DATABASE/abepitopedens.gff: loading...
     DATABASE/abepitopedens.gff: 158 records loaded


     lagbaja:~% bp_genbank2gff.pl –dsn dbi:mysql:HIV1 –adaptor

                dbi::mysql –viral –gb_folder DATABASE/hiv_gff_test_dir/
```

Figure 4-3 (a) The MySQL database underlying the HIV genome browser was created and necessary permissions granted. (b) The database was populated using two Perl scripts named *bp_bulk_load_gff.pl* and *bp_load_gff.pl*. *bp_bulk_load_gff.pl* creates the database tables and populates them. *bp_load_gff.pl* loads data into existing tables of a database. *bp_genbank2gff.pl* converts GenBank format files to GFF format and populates a Bio::DB::GFF database.

4.2.2.2. Sequences and Annotations in the Database

Since HXB2 is used widely and has been annotated extensively, it was used as the key reference sequence in the HIV Genome Browser. HXB2 is the generally accepted HIV-1 reference sequence and is available at GenBank under the accession number K03455. HXB2 is used as a reference genome for all HIV genome sequences because it has been derived from a demonstrably infectious clone (Kuiken et al. 2001). It has been used extensively since 1985 (Leitner et al. 1997) in HIV-1 research and has been annotated completely in terms of genes, coding sequences, proteins, LTRs (long terminal repeats) and other biologically important features. At the Los Alamos HIV Sequence Database coordinates for every annotated feature on HXB2 can be found (http://www.hiv.lanl.gov/content/hiv-db/NUM-HXB2/NUMBERING.html).

T helper cell, CTL and antibody epitopes are available at the Los Alamos Immunology Database for HXB2. Displaying epitope maps is very valuable for the HIV research community as it can give researchers an indication of which regions are more likely to be targeted by an immune response in order to identify possible drug targets. The epitope mappings are available as peptide coordinates and can be obtained as three distinct comma-separated files from (http://www.hiv.lanl.gov/content/immunology/tables/tables.html).

A GFF file was created containing information about the HXB2 genome obtained from both Los Alamos and the GenBank record. The sequence annotations were manually converted to GFF format to ensure accuracy of coordinates. A Perl script was written to extract the necessary information from the comma-separated file containing CTL, antibody and T helper epitopes, mapping the epitope data to HXB2 genome coordinates, and finally creating a GFF file. The complete GFF file contained 2303 lines (see Appendix I for an extract of the GFF file). Separate GFF files were created to load epitope density data. Epitope density files were created by a script *bp_generate_histogram.pl* on GFF files containing separately mapped epitope information.

Whole genome sequences of HIV genomes available in GenBank were retrieved using an SQL query supplied by M. Wilkinson. The GenBank format sequences were converted to GFF format and loaded into the existing HIV1 database with a Perl script *bp_genbank2gff.pl*. These sequences were used to simulate patient sequences and to investigate the impact of a larger database on the performance of the system. Although 284 full-length HIV genome sequences were added to the database no delay in response time was noticed.

Gag sequences from the Viral Diversity Laboratory at the University of Cape Town for a specific patient at different time-points of HIV infection were loaded into the database. The sequences were submitted to ClustalW to perform a multiple sequence alignment. Positions where nucleotides differed between the sequences were noted and a GFF file was created with the various nucleotides occurring at the specific positions. The GFF file was also loaded into the Gbrowse database.

4.2.2.3. The Configuration File

The tracks section of the configuration file contains definition of tracks for the protein-coding genes, frame usage for the different coding regions, the annotated peptides encoded by the different genes, the DNA sequence (at high resolution) and GC content (at low resolution), a display of the three-frame translation displaying either stop and/or start codons (low resolution) or the amino acid sequence (high resolution), the Los Alamos Immunology Database's CTL epitopes relative to HXB2. Other tracks include a track displaying CTL, T helper and antibody epitope density as well as the individual epitopes when zoomed in.

Mutation density in the sequences from UCT can be viewed in a track named *mutation density*. The location of mutations in the genome or the specific nucleotide at a variable position may be viewed in the *mutations* track.

The complete configuration file that was used for the HIV1 database appears in Appendix II. Figures 4-4 to 4-6 are screenshots of the HIV Genome Browser prototype displaying various tracks mentioned in the previous paragraphs.

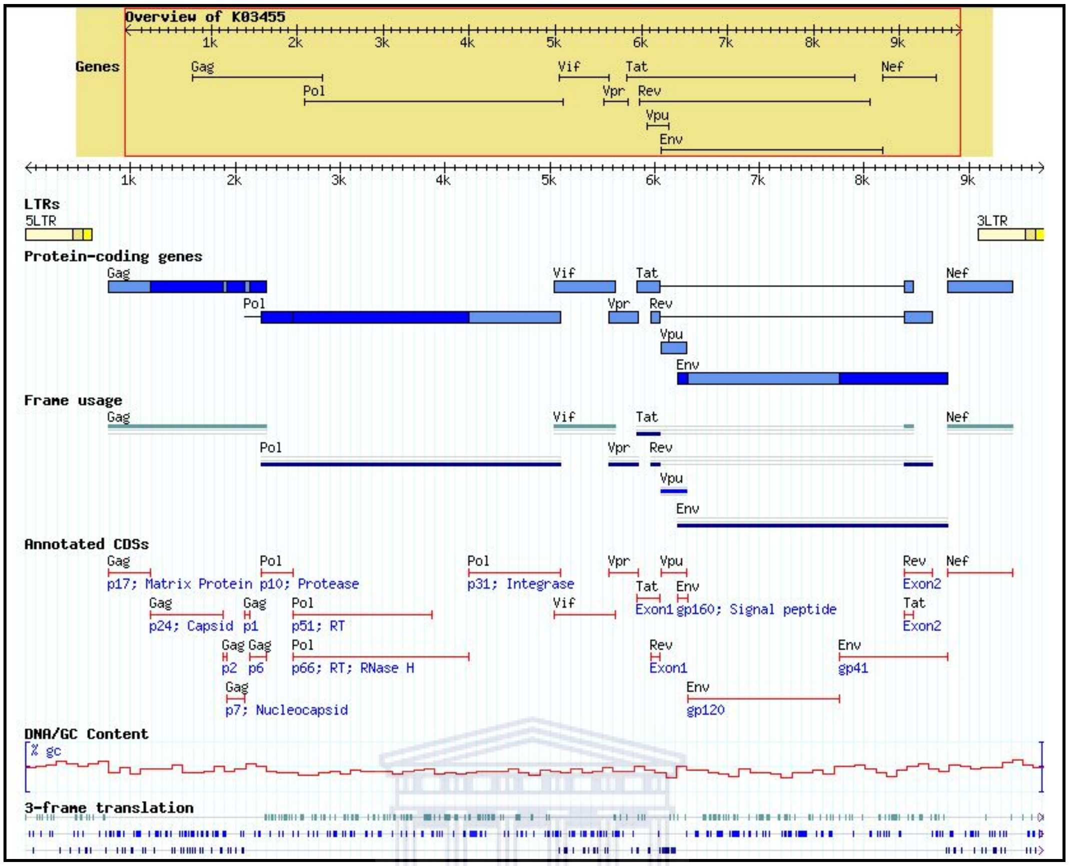


Figure 4-4 A display of general tracks available on the HIV Genome Browser. HXB2 was used as prototype, but it will be possible to display the same data for all sequences loaded into the underlying database.

4.2.2.4. Availability

The HIV Genome Browser can be accessed at http://lagbaja.sanbi.ac.za/cgi-bin/gbrowse/HIV1. To contact the database administrator directly users can send emails to anelda@sanbi.ac.za.

4.3. Example Use Cases

The project for which the HIV Genome Browser prototype was developed (CAPRISA – http://www.caprisa.org) will be generating complete and partial HIV genome sequences as well as epitope (or reactive peptide) sequences for patients at different time points in HIV infection. These sequences will be stored in an integrated molecular database.

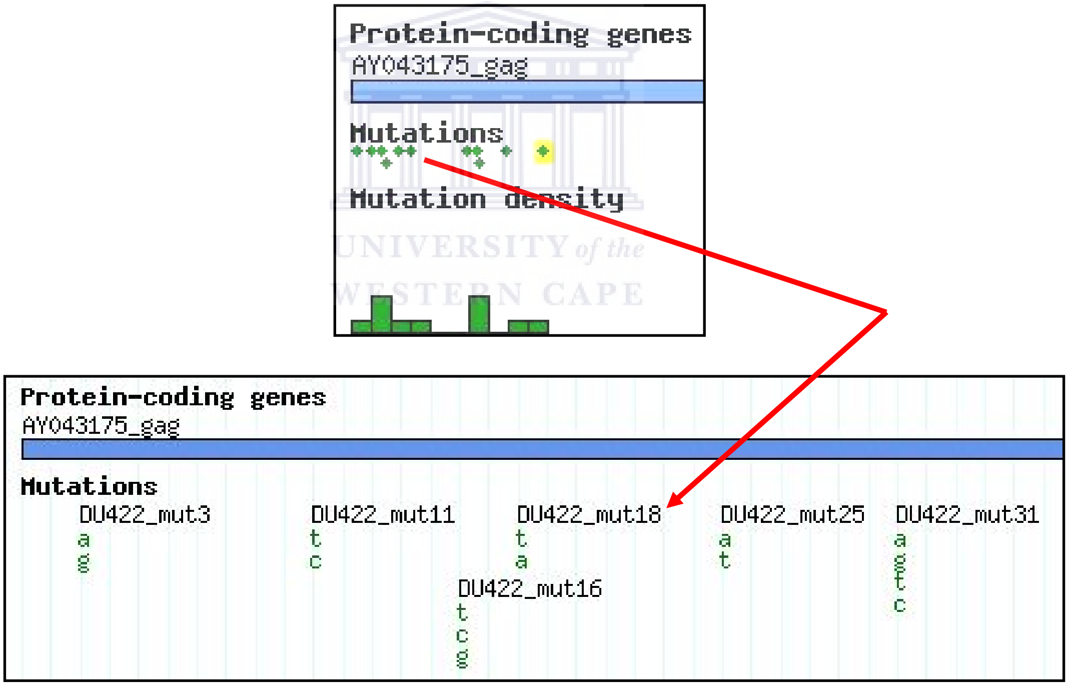
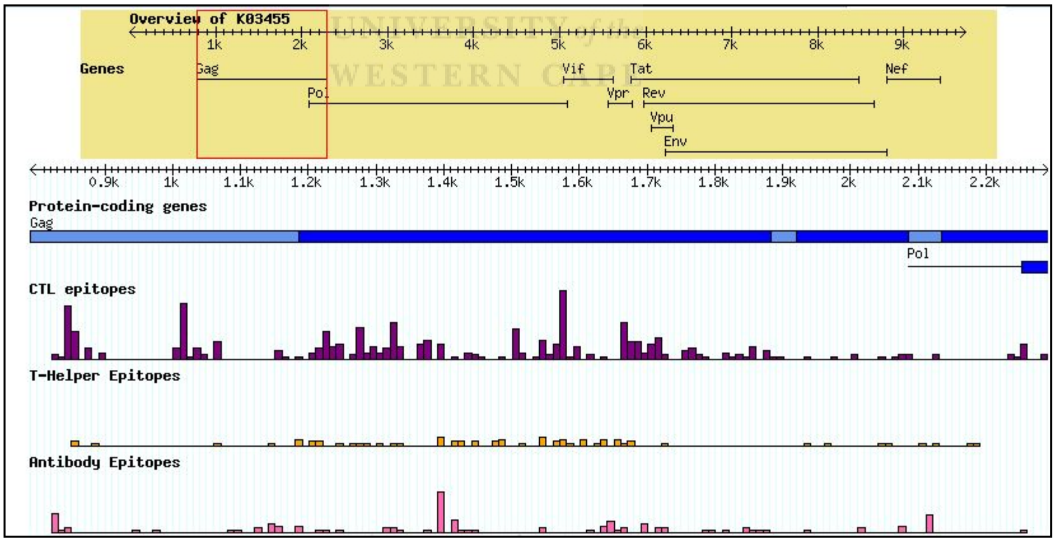


Figure 4-5 Mutations can be viewed as a density profile or specific nucleotides can be displayed. DU422 represents the patient number from whom the sequences were isolated.

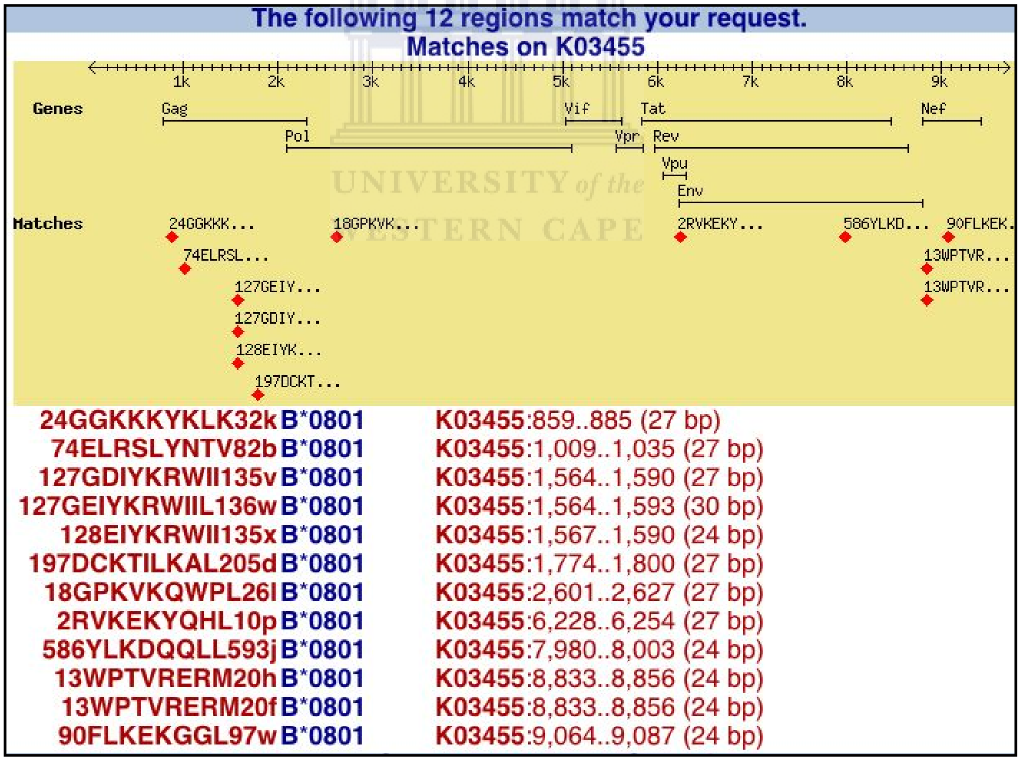
Displaying the accumulation of mutations (Figure 5-1) in the HIV genome sequence infecting each patient over the time course of the infection may yield insight into the evolution of HIV sequences, providing insight into the biology of the virus. Since clinical information will also be available to users, the correlation between sequence mutation and the use of antiretrovirals or other identified selection pressures could be identified.

Regions where large numbers of the population show good epitope responses but that are under low positive selection constraint generally indicates good vaccine candidates (Ernstoff, 2002). Overlaying results of positive selection analyses and epitope density over regions of interest in a graphical way may enhance identification of possible vaccine candidates.



Bench biologists using the genome browser can take advantage of the display of restriction enzyme cut-sites in regions of interest to design PCR experiments and primers.

Where HLA alleles restricting specific viral strains are known (available from the database), users can display an HLA restriction profile across the genome to identify epitopes being restricted by a particlar HLA allele (Figure 4-6).



The following 12 regions match your request.

Matches on K03455

24GGKKKYKLK32k	B*0801	K03455:859..885 (27 bp)
74ELRSLYNTV82b	B*0801	K03455:1,009..1,035 (27 bp)
127GDIYKRWII135v	B*0801	K03455:1,564..1,590 (27 bp)
127GEIYKRWIIL136w	B*0801	K03455:1,564..1,593 (30 bp)
128EIYKRWII135x	B*0801	K03455:1,567..1,590 (24 bp)
197DCKTILKAL205d	B*0801	K03455:1,774..1,800 (27 bp)
18GPKVKQWPL26l	B*0801	K03455:2,601..2,627 (27 bp)
2RVKEKYQHL10p	B*0801	K03455:6,228..6,254 (27 bp)
586YLKDQQLL593j	B*0801	K03455:7,980..8,003 (24 bp)
13WPTVRERM20h	B*0801	K03455:8,833..8,856 (24 bp)
13WPTVRERM20f	B*0801	K03455:8,833..8,856 (24 bp)
90FLKEKGGL97w	B*0801	K03455:9,064..9,087 (24 bp)

Figure 4-6 Researchers can search using an HLA type name (e.g. B*0801) to view the distribution of epitopes in the HIV genome that are restricted by the specific HLA type.

Chapter 5 Conclusions

In this study a set of assessment criteria was developed according to which existing genome browsers were evaluated for their suitability as a starting point to develop a genome browser for the human immunodeficiency virus. Four currently available genome browsers Ensembl, the Generic Genome Browser, the Genome Browser at UCSC and Map Viewer were assessed according to the criteria.

None of the browsers that were evaluated had all the functionality as required by the HIV genome browser's user specifications (as was expected since the existing browsers have been developed for use in eukaryotic genome studies). The time limitation on the project was taken into consideration together with the performance of each of the existing browsers when they were evaluated. GBrowse was chosen as a base for the HIV genome browser prototype. GBrowse was selected as it performed the best in the assessment and was by far the simplest to customize to display HIV genome sequences and annotations. Installation was straightforward. Excellent documentation was available to assist in installation and customization of the browser and the developer community was extremely helpful.

The GBrowse implementation of the HIV genome browser currently displays 284 complete HIV genome sequences together with their annotations as obtained from the GenBank files downloaded from Entrez.

The HIV reference strain, known as HXB2 (GenBank accession number K03455), can also be viewed in the browser. HXB2 has been annotated extensively and for this reason a larger number of annotation tracks can be viewed in relation to the genomic sequence.

The basic annotations available for all genomes available in the database underlying the browser includes genes, LTRs, coding regions, the three-frame translations, DNA sequence, GC content profile, frame usage and restriction enzyme cut sites. Users can also view CTL, T helper and antibody epitope positions for HXB2. Furthermore a representation of the epitope density can be displayed for HXB2. HXB2 was used as a prototype sequence to demonstrate possible features that could be displayed for researchers' proprietary sequences.

All tracks currently displayed in the prototype have been tested on data from previous HIV studies in South Africa and have been demonstrated to participating researchers. The afore-mentioned South African data has not yet been published and could not be made available in this publication.

Mutations mapped onto sequences from a single patient at different timepoints of HIV infection can be visualized for the *gag* gene. These sequences were submitted to ClustalW for a multiple sequence alignment and, for the purpose of this study, positions where nucleotides differ were defined as mutations.

Functionality that was not available through GBrowse included the display of data derived from multiple sequences including entropy graphs, epitope density graphs for different patients, positive selection profiles as calculated from multiple sequence alignments and profiles indicating the number of changes from the reference genome sequence (calculated for each column of a multiple sequence alignment). There are two approaches to be able to view data pertaining to multiple sequences: alterations can be made to the CGI script or plugins can be written. The proposed changes were not added to the GBrowse script nor were any additional plugins written because this fell outside the scope of this project.

Graphs displaying epitope density were implemented for HXB2 using test data in order to prove the usefulness of graphic visualization in the context of the genome and other annotations.

It will be necessary to write a script to convert the results derived from analyses programs such as entropy calculations and positive selection determination into GFF format for addition to the database. The software that will be used by CAPRISA researchers has not been decided on and for this reason scripts to perform these format conversions have not yet been written.

In conclusion, the HIV genome browser prototype developed in this study can be expanded to create a powerful visual data-mining tool to give insight into the biology and epidemiology of HIV and to accelerate the search for HIV vaccine candidates.

Appendix I

```
##gff-version  3
K03455      GenBank     genome      1       9719    .       .       .       ID=K03455
K03455      LANL_seqdb  repeat_region 1     633     .       .       .       ID=5LTR;Name=5'LTR
K03455      LANL_seqdb  long_terminal_repeat 1     455     .       .       .       Parent=5LTR;Name=U5;Note=U5
K03455      LANL_seqdb  long_terminal_repeat 456   551     .       .       .       Parent=5LTR;Name=R;Note=R
K03455      LANL_seqdb  long_terminal_repeat 552   633     .       .       .       Parent=5LTR;Name=U3;Note=U3
K03455      LANL_seqdb  gene    790     2292    .       .       .       ID=Gag;Name=Gag
K03455      LANL_seqdb  CDS     790     1185    .       +       0
      Parent=Gag;Name=p17;Note=p17;Note=Matrix%20Protein
K03455      LANL_seqdb  CDS     1186    1881    .       +       0       Parent=Gag;Name=p24;Note=p24;Note=Capsid
K03455      LANL_seqdb  CDS     1882    1920    .       +       0       Parent=Gag;Name=p2;Note=p2
K03455      LANL_seqdb  CDS     1921    2085    .       +       0       Parent=Gag;Name=p7;Note=p7;Note=Nucleocapsid
K03455      LANL_seqdb  CDS     2086    2133    .       +       0       Parent=Gag;Name=p1;Note=p1
K03455      LANL_seqdb  CDS     2134    2292    .       +       0       Parent=Gag;Name=p6;Note=p6
K03455      LANL_seqdb  gene    2085    5096    .       .       .       ID=Pol;Name=Pol
K03455      LANL_seqdb  CDS     2253    2549    .       +       0       Parent=Pol;Name=p10;Note=p10;Note=Protease
K03455      LANL_seqdb  CDS     2550    3869    .       +       0       Parent=Pol;Name=p51;Note=p51;Note=RT
K03455      LANL_seqdb  CDS     2550    4229    .       +       0
      Parent=Pol;Name=p66;Note=p66;Note=RT;Note=RNase%20H
K03455      LANL_seqdb  CDS     4230    5096    .       +       0       Parent=Pol;Name=p31;Note=p31;Note=Integrase
K03455      LANL_seqdb  gene    5041    5619    .       +       0       ID=Vif;Name=Vif
K03455      LANL_seqdb  CDS     5041    5619    .       +       0       Parent=Vif;Name=Vif
K03455      LANL_seqdb  gene    5559    5847    .       +       0       ID=Vpr;Name=Vpr
K03455      LANL_seqdb  CDS     5559    5847    .       +       0       Parent=Vpr;Name=Vpr
K03455      LANL_seqdb  mRNA 5831   8469    .       .       .       ID=Tat;Name=Tat
K03455      LANL_seqdb  CDS     5831    6045    .       +       0       Parent=Tat;Name=Tat%20Exon%201;Note=Exon1
```

```
K03455	LANL_seqdb	CDS	8379	8469	.	+	1	Parent=Tat;Name=Tat%20Exon%202;Note=Exon2
K03455	LANL_seqdb	mRNA	5970	8653	.	.	.	ID=Rev;Name=Rev
K03455	LANL_seqdb	CDS	5970	6045	.	+	0	Parent=Rev;Name=Rev%20Exon%201;Note=Exon1
K03455	LANL_seqdb	CDS	8379	8653	.	+	2	Parent=Rev;Name=Rev%20Exon%202;Note=Exon2
K03455	LANL_seqdb	gene	6062	6310	.	+	0	ID=Vpu;Name=Vpu
K03455	LANL_seqdb	CDS	6062	6310	.	+	0	Parent=Vpu;Name=Vpu
K03455	LANL_seqdb	gene	6225	8795	.	.	.	ID=Env;Name=Env
K03455	LANL_seqdb	CDS	6225	6311	.	+	0
	Parent=Env;Name=gp160;Note=gp160;Note=Signal%20peptide
K03455	LANL_seqdb	CDS	6312	7757	.	+	0	Parent=Env;Name=gp120;Note=gp120
K03455	LANL_seqdb	CDS	7758	8795	.	+	0	Parent=Env;Name=gp41;Note=gp41
K03455	LANL_seqdb	gene	8797	9417	.	+	0	ID=Nef;Name=Nef
K03455	LANL_seqdb	CDS	8797	9417	.	+	0	Parent=Nef;Name=Nef
K03455	LANL_seqdb	repeat_region	9086	9719	.	.	.	ID=3LTR;Name=3'LTR
K03455	LANL_seqdb	long_terminal_repeat	9086	9540	.	.	.	Parent=3LTR;Name=U5;Note=U5
K03455	LANL_seqdb	long_terminal_repeat	9541	9636	.	.	.	Parent=3LTR;Name=R;Note=R
K03455	LANL_seqdb	long_terminal_repeat	9637	9719	.	.	.	Parent=3LTR;Name=U3;Note=U3
K03455	LANLImmunDB	CTLepi	820	846	.	+	0
	ID=11GELDRWEKI19a;Name=11GELDRWEKI19;prot=p17;hlarestr=B*4002;organism=human
K03455	LANLImmunDB	CTLepi	820	879	.	+	0
	ID=11GELDRWEKIRLRPGGKKKYK30b;Name=11GELDRWEKIRLRPGGKKKYK30;prot=p17;hlarestr=B62;organism=h
uman
K03455	LANLImmunDB	CTLepi	835	879	.	+	0
	ID=16WEKIRLRPGGKKKYK30c;Name=16WEKIRLRPGGKKKYK30;prot=p17;hlarestr=;organism=human
K03455	LANLImmunDB	CTLepi	841	867	.	+	0
	ID=18KIRLRPGGK26d;Name=18KIRLRPGGK26;prot=p17;hlarestr=A3;organism=transgenic%20mouse
K03455	LANLImmunDB	CTLepi	841	867	.	+	0
	ID=18KIRLRPGGK26e;Name=18KIRLRPGGK26;prot=p17;hlarestr=A3;organism=human
K03455	LANLImmunDB	CTLepi	841	867	.	+	0
	ID=18KIRLRPGGK26f;Name=18KIRLRPGGK26;prot=p17;hlarestr=A*0301;organism=human
```

```
K03455	LANLImmunDB	CTLepi	841	867	.	+	0
	ID=18KIRLRPGGK26g;Name=18KIRLRPGGK26;prot=p17;hlarestr=B*0301;organism=human
K03455	LANLImmunDB	CTLepi	841	870	.	+	0
	ID=18KIRLRPGGKK27h;Name=18KIRLRPGGKK27;prot=p17;hlarestr=B27;organism=human
K03455	LANLImmunDB	CTLepi	841	882	.	+	0
	ID=18KIRLRPGGKKKYKL31i;Name=18KIRLRPGGKKKYKL31;prot=p17;hlarestr=A3;organism=human
K03455	LANLImmunDB	CTLepi	841	882	.	+	0
	ID=18KIRLRPGGKKKYKL31j;Name=18KIRLRPGGKKKYKL31;prot=p17;hlarestr=B62;organism=human
K03455	LANLImmunDB	CTLepi	844	870	.	+	0
	ID=19IRLRPGGKK27k;Name=19IRLRPGGKK27;prot=p17;hlarestr=B*2705;organism=scid-hu%20mouse


K03455	LANLImmunDB	ABepi	823	876	.	.	.
	ID=2ELDKWEKIRLRPGGKTLY29e;Name=2ELDKWEKIRLRPGGKTLY29;prot=p17;mab=HyHIV-2;organism=murine
K03455	LANLImmunDB	ABepi	823	876	.	.	.
	ID=2ELDKWEKIRLRPGGKTLY29f;Name=2ELDKWEKIRLRPGGKTLY29;prot=p17;mab=HyHIV-3;organism=murine
K03455	LANLImmunDB	ABepi	823	876	.	.	.
	ID=2ELDKWEKIRLRPGGKTLY29g;Name=2ELDKWEKIRLRPGGKTLY29;prot=p17;mab=HyHIV-5;organism=murine
K03455	LANLImmunDB	ABepi	823	876	.	.	.
	ID=2ELDKWEKIRLRPGGKTLY29h;Name=2ELDKWEKIRLRPGGKTLY29;prot=p17;mab=HyHIV-6;organism=murine
K03455	LANLImmunDB	ABepi	823	879	.	.	.
	ID=2ELDKWEKIRLRPGGKTLY?29i;Name=2ELDKWEKIRLRPGGKTLY?29;prot=p17;mab=HyHIV-4;organism=murine
K03455	LANLImmunDB	ABepi	838	855	.	.	.
	ID=17EKIRLR22j;Name=17EKIRLR22;prot=p17;mab=32/1.24.89;organism=murine
K03455	LANLImmunDB	ABepi	844	900	.	.	.
	ID=19IRLPGGKKKYMLKHVVWAA38k;Name=19IRLPGGKKKYMLKVWAA38;prot=p17;mab=3B10;organism=murine
```

Appendix II

```
[GENERAL]
description     = HIV-1 Test Database
db_adaptor      = Bio::DB::GFF
db_args         = -adaptor      dbi::mysqlopt
                  -dsn          dbi:mysql:database=HIV1:user=www;host=localhost

aggregators     = match
                  coding
                  CTL_profile{CTLepidens}
                  TH_profile{THepidens}
                  AB_profile{ABepidens}
                  LTRs{long_terminal_repeat/repeat_region}
                  MyGenes{CDS/gene}
                  AB{ABepi/ABepitope}
                  CTLs{CTLbin}
                  ABs{ABbin}
                  THs{THbin}
                  mut{bin}

plugins =         Aligner
                  RestrictionAnnotator
                  BatchDumper
                  FastaDumper
                  GFFDumper
                  OligoFinder
                  SequenceDumper

# list of tracks to turn on by default
default features = Transcripts

reference class  = Sequence

# examples to show in the introduction
examples =        K03455:1..9719 gag human B*0801 AF543910:1..403 *mut*
                  DU422_mut* *rev *LTR AY043175:1..100 AY043175:30..99

# "automatic" classes to try when an unqualified identifier is given
automatic classes = Note Name


### HTML TO INSERT AT VARIOUS STRATEGIC LOCATIONS ###
# a footer
footer = <hr><pre>$Id: hiv_test_genbank.conf,v 0.0 2003/11/06 17:32 aboardman Exp
$</pre>
```

```
# what image widths to offer
image widths  = 450 640 800 1024

# default width of detailed view (pixels)
default width = 800

# Web site configuration info
stylesheet  = /gbrowse/gbrowse.css
buttons     = /gbrowse/images/buttons
tmpimages   = /gbrowse/tmp

# max and default segment sizes for detailed view
max segment     = 12000
default segment = 500

# zoom levels
zoom levels    = 100 200 1000 2000 5000 10000 12000

# colors of the overview, detailed map and key
overview bgcolor = #FFFF99
detailed bgcolor = lightpink
key bgcolor      = beige

########################
# Plugin configuration
########################

[Aligner:plugin]
alignable_tracks   = epitopes
upcase_tracks      = CDS
upcase_default     = CDS

########################
# Default glyph settings
########################

[TRACK DEFAULTS]
glyph         = generic
height        = 10
bgcolor       = lightgrey
fgcolor       = black
font2color    = blue
label density = 100
bump density  = 100
# where to link to when user clicks in detailed view
link          = AUTO
```

```
################## TRACK CONFIGURATION ####################
# the remainder of the sections configure individual tracks
###########################################################
[Genes:overview]
feature         = MyGenes
glyph           = span
height          = 5
description     = 0
label           = 1
key             = Genes

 [Transcripts]
feature         = MyGenes
glyph           = graded_segments
bgcolor         =  sub  {
                    my $region = shift;
                    return "#0000FF" if $region->attributes('Name') eq 'p10';
                    return "#6699FF" if $region->attributes('Name') eq 'p51';
                    return "#0000FF" if $region->attributes('Name') eq 'p66';
                    return "#6699FF" if $region->attributes('Name') eq 'p31';
                    return "#0000FF" if $region->attributes('Name') eq 'gp160';
                    return "#6699FF" if $region->attributes('Name') eq 'gp120';
                    return "#0000FF" if $region->attributes('Name') eq 'gp41';
                    return "#6699FF" if $region->attributes('Name') eq 'p17';
                    return "#0000FF" if $region->attributes('Name') eq 'p24';
                    return "#6699FF" if $region->attributes('Name') eq 'p2';
                    return "#0000FF" if $region->attributes('Name') eq 'p7';
                    return "#6699FF" if $region->attributes('Name') eq 'p1';
                    return "#0000FF" if $region->attributes('Name') eq 'p6';
                    return "#6699FF";
          }
fgcolor= black
description     = 0
key             = Protein-coding genes
citation        = Showing introns and exons in the case of Rev and Tat.
title           = $start $end
```

References

Abbas AK & Lichtman AH (eds). (2001). *Basic Immunology: Functions and disorders of the immune system.* WB Saunders Company, Curtis Centre, Philadelphia, USA.

Bergh S & Cole ST. MycDB: An integrated mycobacterial database. *Mol. Microbiol*. 12 (1994): 517-534.

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez X.M, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M & Hubbard T. Ensembl 2004. *Nucleic Acids Res*. 32 Database issue (2004): D468-470.

Birney E, Clamp M & Hubbard T. Databases and tools for browsing genomes. *Annu.Rev.Genomics Hum.Genet*. 3 (2002): 293-310.

Chen, JY & Carlis, JV. Genomic Data Modeling. *Information Systems.* 28 (2002): 287-310.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S & Botstein D. SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* 26.1 (1998): 73-80.

Chi EH, Barry P, Shoop E, Carlis JV, Retzel E & Riedl J. Visualization of biological sequence similarity search results. *Proceedings of the 6th IEEE Visualization Conference*. (1995): 44-51.

Choudhuri S. The path from nuclein to human genome: A brief history of DNA with a note on human genome sequencing and its impact on future research in biology. *Bulletin of Science Technology Society*. 23 (2003): 360-367.

Crandall KA (ed). (1999). *The evolution of HIV*. The John Hopkins University Press, Maryland, USA.

De Oliveira T, Miller R, Tarin M & Cassol S. An integrated genetic data environment (GDE)-based LINUX interface for analysis of HIV-1 and other microbial sequences. *Bioinformatics*. 19.1 (2003):153-154.

Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4:R60 (2003): 1-11.

Dowell RD, Jokerst RM, Day A, Eddy SR & Stein L. The Distributed Annotation System. *BMC Bioinformatics.* 2.7 (2001): 1-7.

Ernstoff EA. (2002). *HIV subtype C diversity: Analysis of the relationship of sequence diversity to proposed epitope locations.* Unpublished MSc thesis. Faculty of Science, University of the Western Cape, South Africa.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J, Dougherty BA, Merrick JM, McKenney K, Sutton GG, FitzHugh W, Fields CA, Gocayne JD, Scott JD, Shirley R, Liu LI, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback T, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NS, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. Whole-genome random sequencing and assembly of Haemophilus influenzae. *Science*. 269.5223 (1995): 496-512.

Frost SD, Gunthard HF, Wong JK, Havlir D, Richman DD & Leigh Brown AJ. Evidence for positive selection driving the evolution of HIV-1 env under potent antiviral therapy. *Virology*. 5.284 (2001):250-258.

Globus A & Uselton S. Evaluation of visualization software. *Computer Graphics*. 29.2 (1995): 41-44.

Helt GA, Lewis S, Loraine AE & Rubin GM. BioViews: Java-based tools for genomic data visualization. *Genome Res*. 8.3 (1998): 291-305.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I & Clamp M. The Ensembl genome database project. *Nucleic Acids Res*. 30.1 (2002): 38-41.

Kantor R, Machekano R, Gonzales MJ, Dupnik K, Schapiro JM & Shafer RW. Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database: an expanded data model integrating natural language text and sequence analysis programs. *Nucleic Acids Res*. 29.1 (2001): 296-299.

Kashuk C, SenGupta S, Eichler E & Chakravarti A. viewGene: A graphical tool for polymorphism visualization and characterization. *Genome Res.* 12 (2001): 333-338.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T & Birney E. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*. 14.1 (2004): 160-169.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM & Haussler D. The human genome browser at UCSC. *Genome Res*. 12.6 (2002): 996-1006.

Klausner RD, Fauci AS, Corey L, Nabel GJ, Gayle H, Berkley S, Haynes BF, Baltimore D, Collins C, Douglas RG, Esparza J, Francis DP, Ganguly NK, Gerberding JL, Johnston MI, Kazatchkine MD, McMichael AJ, Makgoba MW, Pantaleo G, Piot P, Shao Y, Tramont E, Varmus H & Wasserheit JN. Enhanced: The Need for a Global HIV Vaccine Enterprise. *Science.* 300.5628 (2003): 2036-2039.

Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S &Korber B (eds). *HIV Sequence Compendium 2001.* Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Laird SP, Wong JSK, Schaller WJ, Erickson BJ & De Groen PC. Design and implementation of an Internet-based medical image viewing system. *Journal of Systems and Software.* 66 (2003): 167-181.

Leader DP. BugView: a browser for comparing genomes. *Bioinformatics*. 20.1 (2004): 129-130.

Lehohla PJ. *Mid-year population estimates, South Africa 2004. Statistical Release* P0302. (2004). Statistics South Africa, Pretoria, South Africa.

Leitner T, Korber B, Robertson D, Gao F & Hahn B. (1997). *Updated Proposal of Reference Sequences of HIV-1 Genetic Subtypes. pp. III-19-24 in Human Retroviruses and AIDS 1997.* Korber,B, Hahn,B, Foley,B, Mellors,JW, Leitner,T, Myers,G, McCutchan,F & Kuiken,CL (eds). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smith CD, Tupy JL, Rubin GM, Misra S, Mungall CJ &Clamp ME. Apollo: a sequence annotation editor. *Genome Biol.* 3 (2002): research0082.1-0082.14.

Loraine AE & Helt GA. Visualizing the Genome: techniques for displaying human genome data. *BMC Bioinformatics*. 3.19 (2002): 1-8.

Matthiessen MW. Affordable biocomputing for everyone: using the Internet, freeware and open-source software. *Trends Biochem.Sci.* 27.11 (2002): 586-588.

McMichael A, Mwau M & Hanke T. HIV T cell vaccines, the importance of clades. *Vaccine*. 20.15 (2002): 1918-1921.

Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris N, Marshall B, Shu S, Kaminker JS, Prochnik SE, Smith CD, Smith E, Tupy JL, Wiel C, Rubin GM & Lewis SE. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol*. 3.12 (2002): RESEARCH0081.

Navarro JD, Niranjan V, Peri S, Jonnalagadda CK & Pandey A. From biological databases to platforms for biomedical discovery. *Trends Biotechnol*. 21.6 (2003): 263-268.

Pollet N, Schmidt HA, Gawantka V, Vingron M & Niehrs C. Axeldb: a Xenopus laevis database focusing on gene expression. *Nucleic Acids Res.* 28.1 (2000): 139-140.

Pomerantz RJ. HIV: cross-talk and viral reservoirs. *Nature.* 424 (2003): 136-137.

Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J & Shafer RW. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. 31.1 (2003): 298-303.

Robinson AJ & Flores TP. Novel techniques for visualising biological information. *Proc.Int.Conf.Intell.Syst.Mol.Biol*. 5 (1997): 241-249.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M & Barrell B. Artemis: sequence visualization and annotation. *Bioinformatics* 16.10 (2000): 944-945.

Shafer RW, Jung DR, Betts BJ, Xi Y & Gonzales MJ. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. 28.1 (2000): 346-348.

Shafer RW, Jung DR & Betts BJ. Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nat.Med.* 6.11 (2000): 1290-1292.

Shafer RW, Stevenson D & Chan B. Human Immunodeficiency Virus Reverse Transcriptase and Protease Sequence Database. *Nucleic Acids Res*. 27.1 (1999): 348-352.

Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel EW, Rubin EM, Hamann B & Dubchak I. Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics.* 20.5 (2004): 636-643.

Shneiderman B. Response time and display rate in human performance with computers. *ACM Computing Surveys.* 16.3 (1984): 265-285.

Stein LD, Cartinhour S, Thierry-Mieg D & Thierry-Mieg J. JADE: an approach for interconnecting bioinformatics databases. *Gene*. 209.1-2 (1998): GC39-GC43.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A & Lewis S. The generic genome browser: a building block for a model organism system database. *Genome Res*. 12.10 (2002): 1599-1610.

Stein LD & Thierry-Mieg J. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACeDB databases. *Genome Res*. 8 (1998): 1308-1315.

Stein LD & Thierry-Mieg J. AceDB: a genome database management system. *Computing in Science and Engineering*. 1.3 (1999): 44-52.

Suzuki Y & Gojobori T. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol*. 16.10 (1999): 1315-1328.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA & Wagner L. Database resources of the National Center for Biotechnology. *Nucleic Acids Res*. 31.1 (2003): 28-33.

Wilkinson MD, Block D & Crosby WL. Genquire: genome annotation browser/editor. *Bioinformatics*. 18.10 (2002): 1398-1399.