The Molecular Integration Database (MID) for democratizing HIV data

Heikki Lehväslaiho

29 May 2007, Nairobi

Bioinformatics for Africa

Classical pathogen biology

- clinical aspects of the host
- physiology
- immunology
- biochemistry
- genetics

medical bioinformatics

Pathogen biology today

- **genomics**
  - especially when applied to small genomes!
- biochemistry
- immunology
- clinical aspects of the host
- physiology

Genomics paradigm

- genome variation is central
  - both host and pathogen
- time series of samples
- tools are freely available and timely distributed
- data are freely available and timely distributed

What is needed?

- database engine
- query interface
- visualisation

freely available tools

Prerequisites for tools in MID

- Open source
- Under active development
- Web interface
  - build if needed
- Derived work has to be distributable

What is needed?

- database engine
- query interface
- visualisation

freely available tools

- (fast) internet connection
- knowledge on bioinformatics
- translational skills

capacity building

Multidimensional data

- researches not in the same location
  - communication is difficult
- researches are specialist in different fields
  - "translation of knowledge" is needed

How does this work?

- Consult researchers
- Create a central service with tools
- Consult researchers
- Create better tools
- Consult researchers
- Distribute new tools
- Consult researches

Purpose of the MID

- Integrate HIV biomedical information created by several laboratories into one query enabled interface
  - A big step up from spreadsheets
  - Central location for all ***exchange*** information
    - backups
    - versioning
    - audit trail

Not a LIMS!

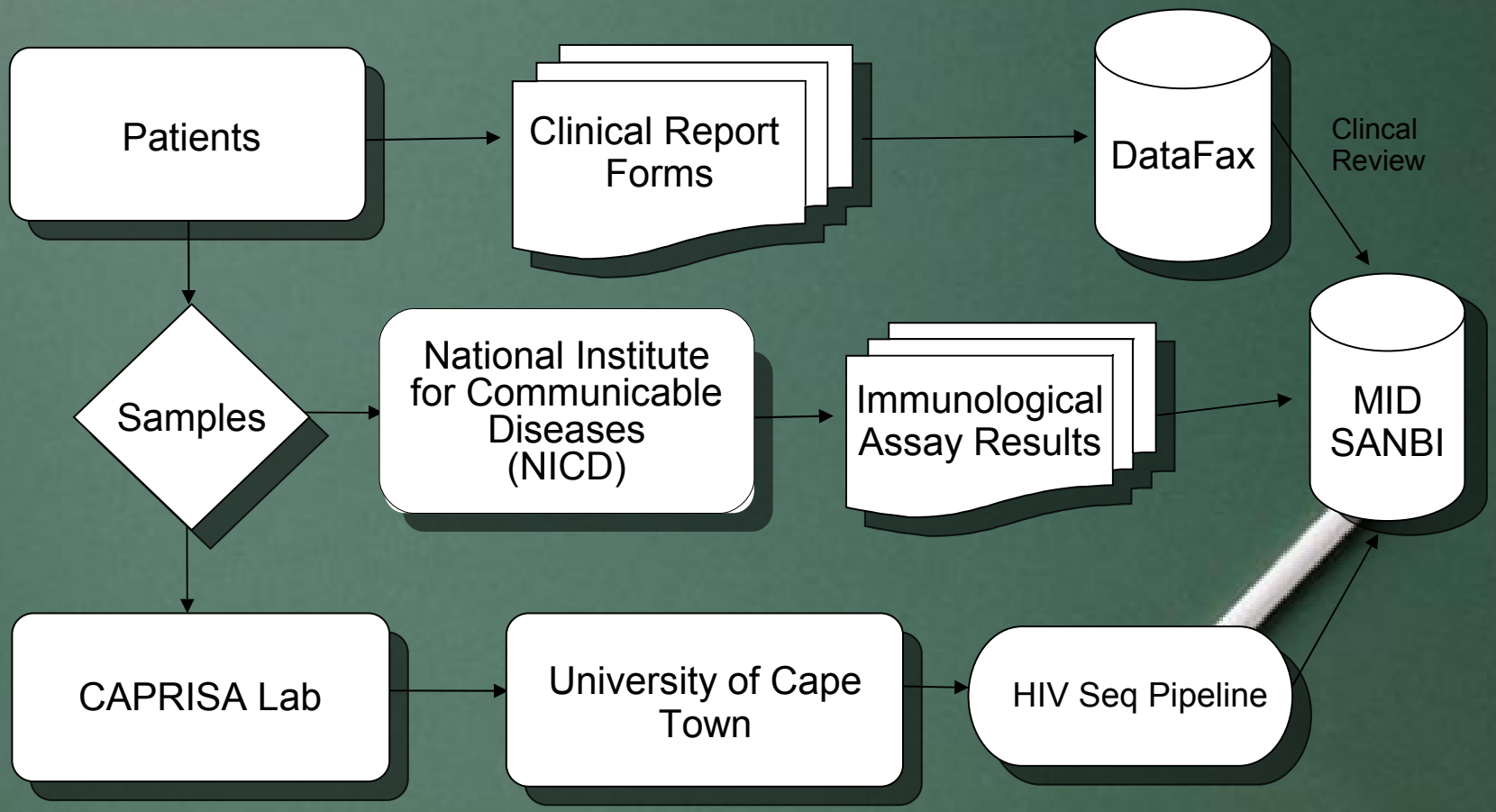
CAPRISA

CENTRE FOR THE AIDS PROGRAMME OF RESEARCH IN SOUTH AFRICA

CAPRISA IS A UNAIDS COLLABORATING CENTRE FOR HIV PREVENTION RESEARCH

- Elucidate HIV pathogenesis and immune escape that influence the set point in heterosexually acquired HIV subtype C infection
- Cohort of female sex workers in KwaZulu-Natal, participants in a Phase II/IIb microbicide trial in Durban, cohorts in Vulindlela, KZN

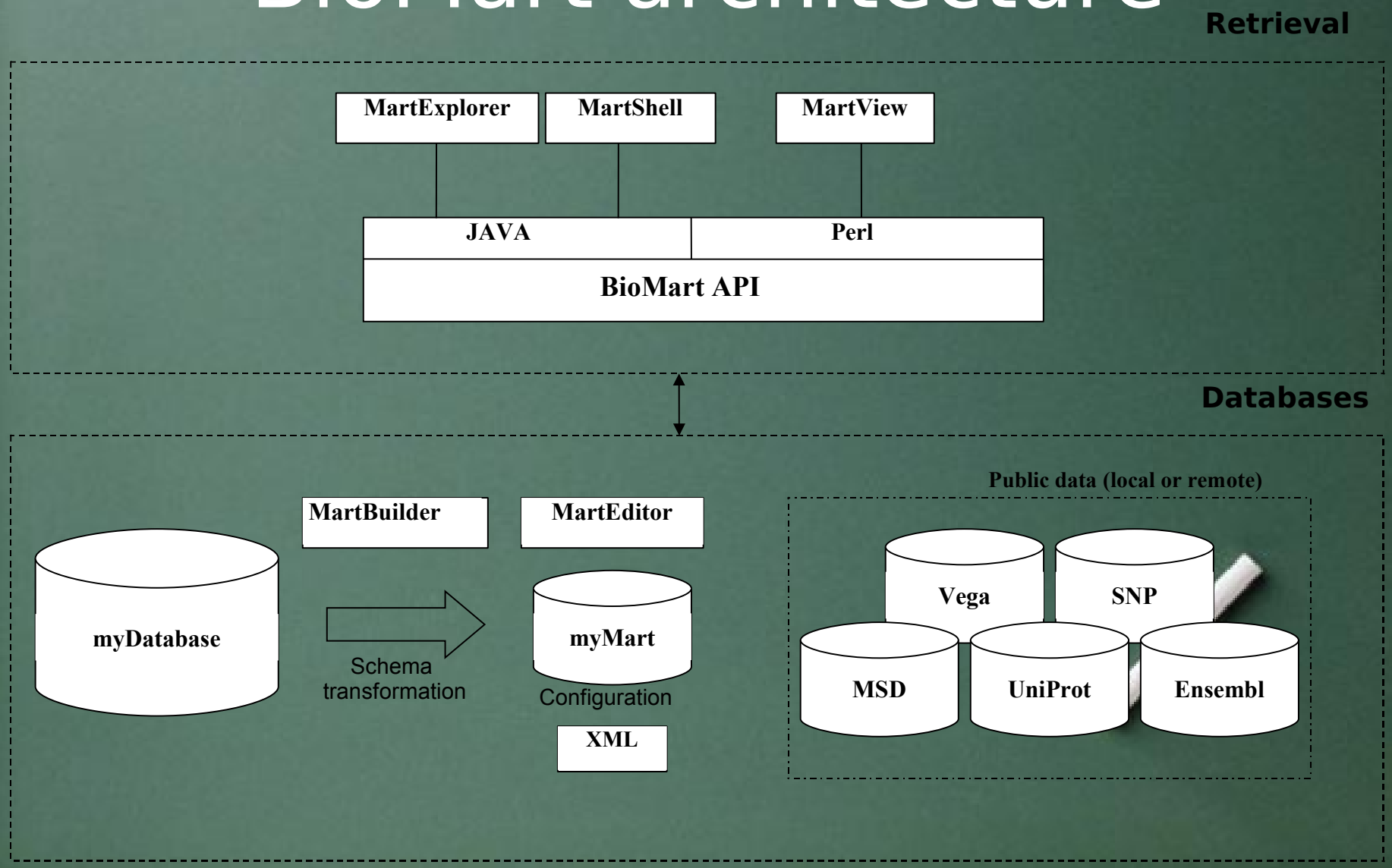
CAPRISA data collection



The databases

- production database
  - MySQL, PostgressQL
  - modified by the maintainer only (security)
  - generic tables to minimize changes
- query database
  - derived from the production db
  - read-only by the authorised users
  - BioMart: www.biomart.org

BioMart architecture



Choose the data set

New | XML | Help | Count | Results

» **Dataset:**
mid
» **Attributes**
[None selected]
» **Filters**
[None selected]

» **Dataset:**
[None Selected]

Database: MID Biomart beta (BioMart3 PostgreSQL Database)

Dataset: mid

Using MartView

1. Choose **Dataset** above
2. Click **Attributes** and make your selection in this panel
3. Click **Results** in the top panel

You can further refine your query by including **Filters** and/or additional **Dataset**

Mini Tutorial

biomart version 0.5

Choose the output

New | XML | Help | Count | Results

» **Dataset:**
mid
» **Attributes**
Participant id
Phase
Visit
Interim visit
Months post infection
» **Filters**
[None selected]

» **Dataset:**
[None Selected]

⊟ Participant timepoint

participant

- [ ] Study id
- [ ] Location
- [x] Participant id

participant timepoint

- [ ] S I p p v i
- [x] Phase
- [x] Visit
- [x] Interim visit
- [ ] Year
- [ ] Month
- [ ] Day

⊟ Period post infection

period post infection

- [ ] Period post infection exists
- [ ] Days post infection
- [ ] Weeks post infection
- [x] Months post infection
- [ ] Infection year
- [ ] Infection month
- [ ] Infection day

⊞ HIV sequence

biomart version 0.5

Select the filters

New | XML | Help | Count | Results

» **Dataset:**
mid
» **Attributes**
Participant id
Phase
Visit
Interim visit
Months post infection
VL log
CD4
Gene region
» **Filters**
Gene region : Gag
Accession value : [ID-list specified]

» **Dataset:**
[None Selected]

⊞ Period post infection

⊞ Assay data existance

⊟ HIV Sequence

☑ HIV Sequence

Subtype: C

Gene region: Complete genome, Env, Gag, Nef

Accession value: Browse...

biomart version 0.5

View the output

New | XML | Help | Count | Results

» **Dataset:** 5 / 270 Entries
mid
» **Attributes**
Participant id
Phase
Visit
Interim visit
Months post infection
VL log
CD4
Gene region
» **Filters**
Gene region : Gag
Accession value : [ID-list specified]

» **Dataset:**
[None Selected]

Display maximum 20 rows as HTML
Export all results to File Go

Participant id	Phase	Visit	Interim visit	Months post infection	VL log	CD4	Gene region
0177	3	09	0	6	4		
0225	3	10	0	6	4		
0037	3	08	0		4		
0225	2	01	0		4		
0268	2	05	0		3		
0217	2	03	0	3	4		
0225	2	06	0	3	4		
0037	2	05	0	3	4		

Unpublished data

biomart version 0.5

Other parts of MID

- Summary
  - patient data progress monitoring
- glossary
- data entry programs
  - sequence pipeline
    - for sequence assembly and quality control
- visualisation
- bug reporting

CAPRISA data summary

Reports [tools.caprisa.org]

Participant

- 0008
- 0030
- 0037
- 0040
- 0045
- 0061
- 0063
- 0065
- 0069
- 0084
- 0085
- 0088
- 0129
- 0136
- 0137

MID Summary report

This report was generated on: 2007-03-12T17:31:59.052+02:00

PTID	autologous neutralization assay	clinical report	heterologous neutralization serum assay	heterologous neutralization virus assay	hiv sequence	hla typing assay	initial elispot results assay	monoclonal antibody assay	period post infection	phenotype assay	single peptide confirmation assay	viral isolation assay
0008												
0030												
0037												
0040												
0045												
0061												
0063												
0065												
0069												
0084												
0085												
0088												
0129												

unpublished data

CAPRISA HIV Glossary

PTID (Period post infection data)

The first 2 digits in this field consist of the the location code (example 10 stands for Vulindlela), digits at position 3 to 6 represent the participant's unique number.Example: 100045

Visit code (Period post infection data)

The phase, visit and interim visit are contained within this field's 4 characters values. Character at position 1 (first) is the phase, characters at position 2 and 3 represent the visit and character at position 4 is the interim visit.

Visit date (Period post infection data)

The date the visit was made by the participant or when sample was drawn from participant. The date the participant made the provided phase, visit, and interim visit. provided in field 1.

Days post infection (Period post infection data)

The number of days this participant's visit since infection date.

Weeks post infection (Period post infection data)

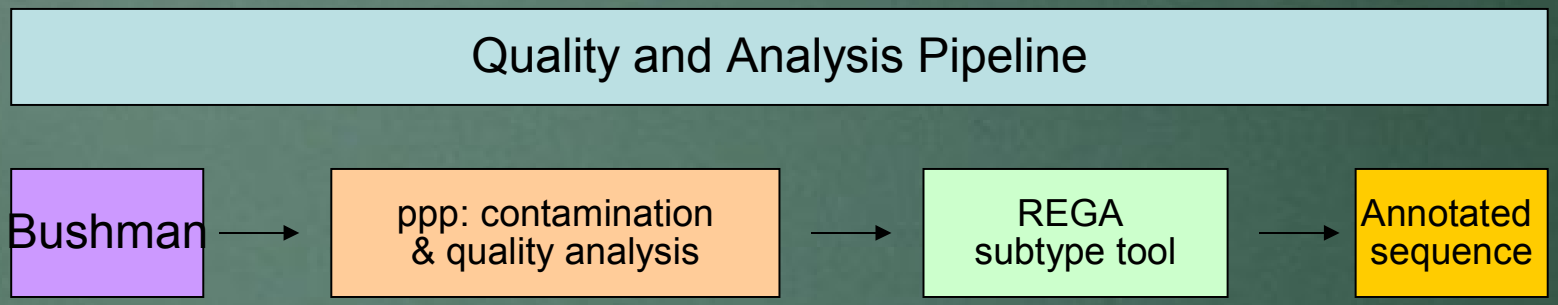
The number of weeks elapsed since the date of infection

Date of infection (Period post infection data)

the date the infection occurred

Sequence pipeline

- One program is now split into three:



  - Bushman: assembly manager
    - general purpose tool
    - installable locally to save bandwidth
  - ppp: HIV quality control pipeline
  - REGA: HIV subtyping tool

Bushman, beta version

South African National Bioinformatics Institute

Assembly Manager

BUSHMAN

HOME

PROJECT MANAGEMENT

- Create new project
- Update existing project
- Remove project

ASSEMBLY AND REPORT

- Assemble chromatograms
- Generate quality report

OUTPUT

- Download fasta sequence
- Read quality report

NEWS

VIEW AND REPORT BUGS

Create new Project

User Name*
guest

Project Name*
mytest

Date Created
Fri May 25 11:39:40 2007

Upload Zipfile
/home/ruby/prj/BushMan/ Browse...

Assemble and Report
Yes No

Keep Backup of Upload file
Yes No

Reset Create

Summary

Project

Name: None

Status: Unselected

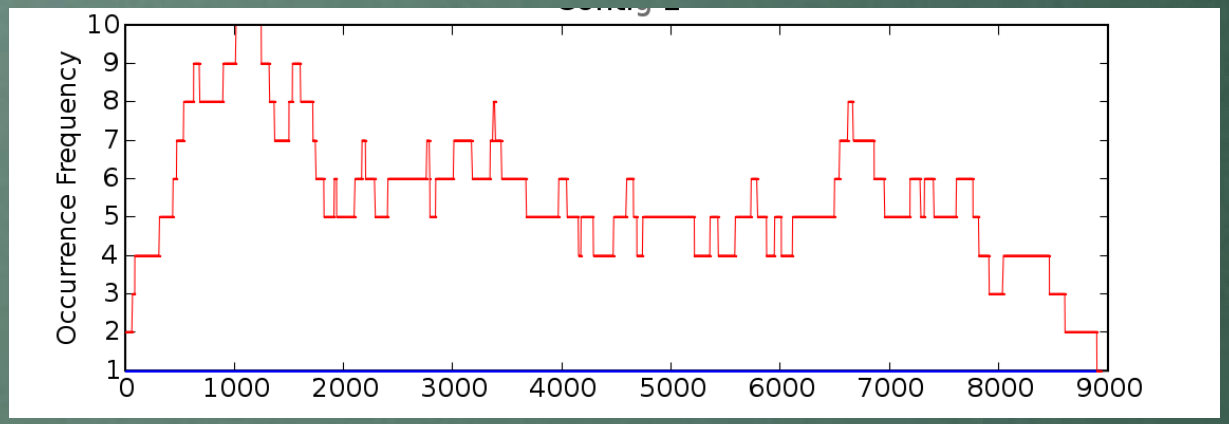
Process

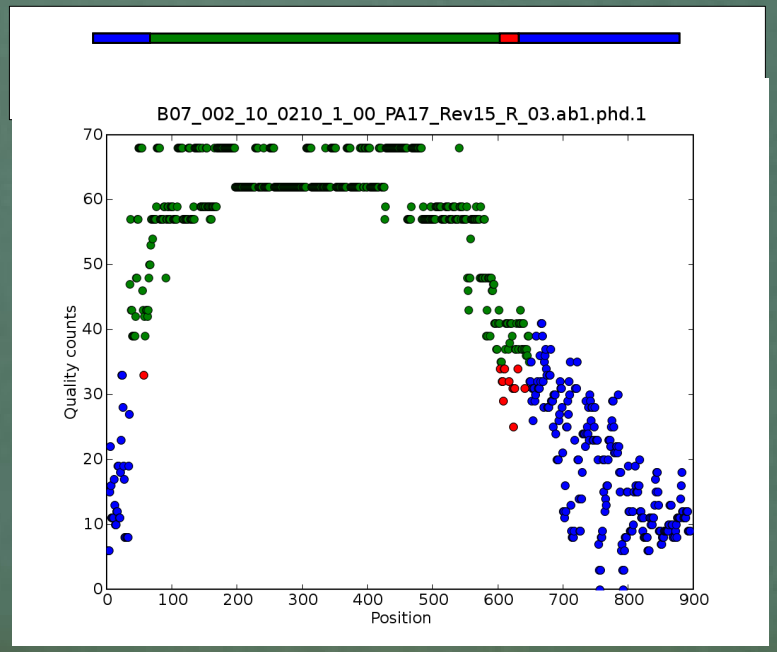
Status: Unassembled

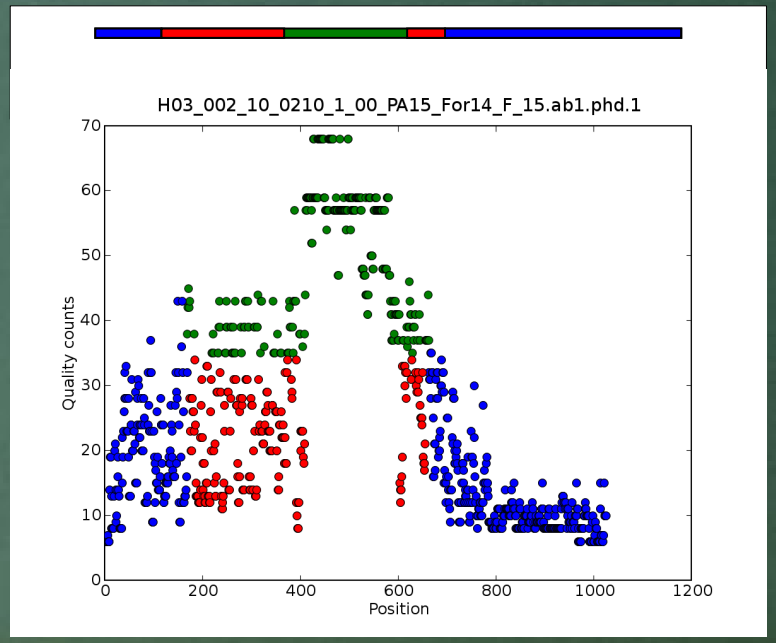
Report

Status: Unavailable

BushMan report







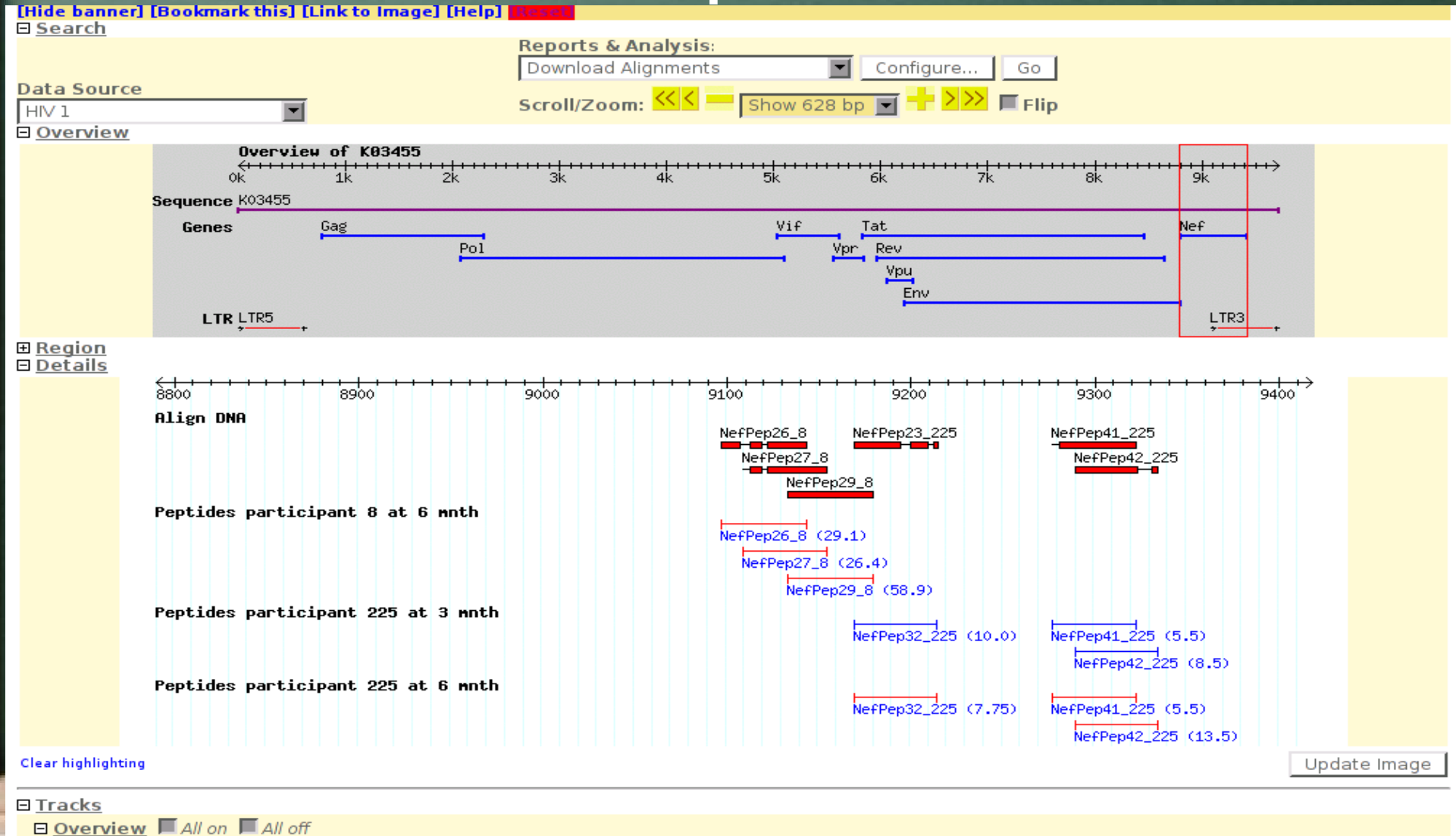
Public Perl Pipeline for HIV

- Reads in the consensus sequence
- Outputs annotation in XML
- fast and easily configurable
- Check for
  1. Contaminants
  2. Location and orientation in the reference HXBX sequence
  3. Protein sequence changes

Visualisation

- Gbrowse genome viewer
  - next: seamless linking between BioMart and GBrowse
- General purpose plotting tool

Low resolution of view of NEF peptide mismatches over time in two patients



Challenges

- low bandwidth
- Complexity of data and user interface
  - GUI design
  - sparsely populated huge data matrix
  - subtle errors -> data cleaning

Contributors

SANBI

- Allan Kamau (database programming)
- Ruby van Rooyen (interface programming)
- Adam Dawe (content testing, user)
- Kavisha Ramdayal (testing, glossary)
- Alan Powell (interface programming)
- Anelda Boardman (laboratory coordination)
- Tulio de Oliviera (HIV evolution, testing)
- Heikki Lehväslaiho (project manager, programming)
- Win Hide (project co-ordinator)

CAPRISA

- Salim Abdool Karim (CAPRISA)
- Koleka Mlisana (CAPRISA)
- Lynn Morris (NICD)
- Clive Gray (NICD)
- Carolyn Williamson (UCT)