Challenges in providing a genome browser for large scale projects

Cancer Genome Project, Sulston Laboratories, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, England

Abstract:

Within the Cancer Genome Project (CGP) we hold data for ~18,000 sequencing experiments. Up to now we have attempted to keep all data available within a standard GBrowse instance but we are now starting to reach the limit of what is feasible to present as individual datasources.

CGP is now working towards generating popup instances of JBrowse which are configured on a per-user basis based on the data that they need for their individual analysis.

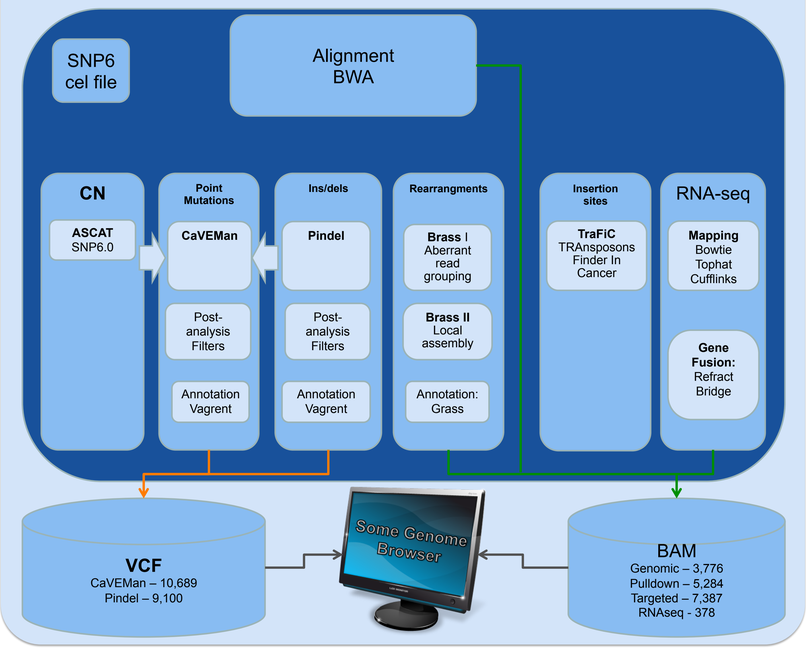
Analysis Pipeline

CGP have developed a mapping and analysis pipeline which is tightly integrated into internal LIMS and Farm architectures.

This allows for easy web-based administration of the system as well as providing submission tools suitable for our scientists.

The system includes analysis for the following:

- Copy Number Variation – SNP6.0 or Paired NGS BAM.
- Point mutations (CaVEMan - Cancer Variants through Expectation Maximisation).
- Insertion/Deletion - Pindel.
- Rearrangements (BRASS - BReakpoint ASSembly).
- TraFiC - Transposon mediated insertions.
- RNA specific analysis.



GBrowse

CGP have been using GBrowse since late 2009.

The initial implementation only supported:

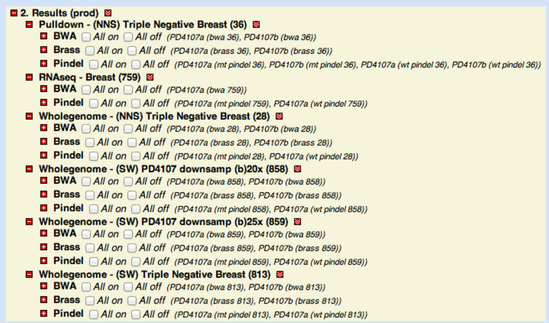
- Human GRCh37
- BWA BAM alignment

Data was relatively simply grouped:

- A particular sequencing type (genomic/pulldown)
- Grouping of samples by tissue or grant.

As our analysis has progressed more complex configurations of datasource have become required to support the following:

- Multiple species/builds
  - 9 distinct species
  - 16 species/builds
- Comparison of 'individual/patient' data from multiple sequencing types or alternative mapping parameters.
- Inclusion of alignments from:
  - RNA seq
  - Rearrangement analysis
  - Ins/del analysis



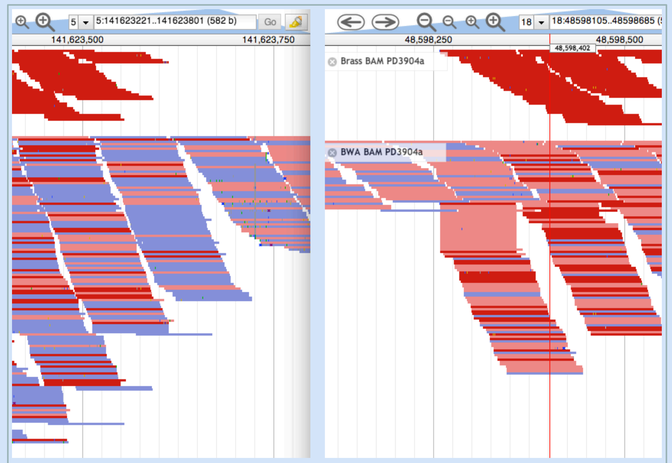
- Remote GFF3 result tracks via custom tracks
  - Loss of Safe::World results in inconsistent display so not configured by default.
  - Provided via web-service conversion from VCF->GFF3
- 26,667 datasources
  - 11,095 hidden but accessible via URL

JBrowse

CGP are in to process of migrating to JBrowse for day to day use.

Primary features driving change:

- No need for database backend and associated resources.
- Much simpler setup and administration of server.
- Very low resource requirement.
- Direct support for Tabix indexed VCF.



Example of clear rearrangement breakpoint. Top tracks show reads spanning breakpoint, lower track shows corresponding drop in mapped reads.

Main item preventing complete migration:

- No ability to save image via scripting
  - Attempted use of phantomJs but no image generated

Planned contributions:

- Generic 'grid to config' script that can take tsv listing to generate a full configuration for JBrowse based on simple config.
- Lightweight 'user' config CGI and scripts to augment any predefined configuration with data combinations defined by a user.

CGP:

PI: Peter Campbell
IT Leads: Adam Butler, Jon Teague
Principal Bioinformatician: Keiran Raine (kr2@sanger.ac.uk)
Development team: Serge Dronov, Jon Hinton, David Jones, Catherine Leroy, Andrew Menzies, Lucy Stebbings