Use of GBrowse 1.999 in the Cancer Genome Project

Keiran Raine
Cancer Genome Project
Wellcome Trust Sanger Institute
kr2@sanger.ac.uk

Overview

- A brief history of the CGP
- COSMIC - DAS vs. GFF3
- Next Gen pitfalls
- Exome ~ 5GB BAM x2
- Wholegenome ~ 100GB BAM x2

A brief history of CGP

- Formed in 2000
- Main screening methods:
  - Hetroduplex
  - Capillary sequencing
  - Illumina Next Gen
- In 2004 COSMIC became a public resource

A brief history of CGP

2000	Hetroduplex screening with capillary confirmation	2 IT staff
2003/4	Capillary sequencing	8 IT staff CSA developed
2000/10	Next Gen with capillary confirmation	12 IT staff

One day in a meeting…

- How are we going to view next gen seq data?
  - 6-12 months development
    - No Chance!
- Go think about it, can we have a solution within the month?
- GBrowse2 won!

COSMIC

- Catalogue Of Somatic Mutations In Cancer
- Database available to all as download
- Need your own front end
- DAS currently being updated to latest spec to maintain Ensembl draw functionallity

COSMIC - DAS vs. GFF3

- New DAS works in GBrowse2 (ish)
  - Not searchable
- Internally showing confirmed variations via GFF3 in GBrowse2
  - Trivial to provide GFF3 exports
  - Will probably include config for GBrowse2 as well
- KRAS data in COSMIC

Some Next Gen bloopers

- Don’t get human ref seq from Ensembl to use for mapping
  - Y has PAR regions masked to ‘n’
  - Can artificially make reads map uniquely
- Select your MT sequence carefully
  - Genbank has recently been updated to show the correct seq
  - Previously was sequence for a small village in Africa (will not make it into Ensembl for a while due to build cycles)

Exome data

- Targeted to exons
- High depth, low cost
- Quick (few lanes needed)
- Works well

Wholegenome data

- Reads are random but size selected
- Can detect rearrangements
- Low depth, high cost
- Takes much longer
- Gbrowse needs to work with 2x100GB files…. It does it well

Acknowledgments

- Sanger Systems
- CancerIT (CGP it group)
- Most frequent respondents to my problems:
  - Lincoln Stein
  - Scott Cain
- Thanks to all on the Gbrowse mailing list for putting up with me.