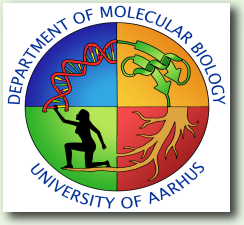
Next generation sequencing

Data visualization

Stig Uggerhøj Andersen, PhD

Department of Molecular Biology

University of Aarhus



Lecture overview

- Data formats
- Visualization software
  - IGV http://www.broadinstitute.org/igv/
  - Galaxy https://main.g2.bx.psu.edu/
  - UCSC browser http://genome.ucsc.edu/
  - Gbrowse http://gmod.org/wiki/Gbrowse
  - Jbrowse http://jbrowse.org/
  - IGB http://bioviz.org/igb/

Good data analysis practices

- Use **data summaries** and **visualization** for sanity checks of each analysis step
- Take advantage of **biological insight** whenever possible
- Improve analysis by evaluation of **representative examples**
- Document and automate **analysis workflows** to ensure efficiency and reproducibility

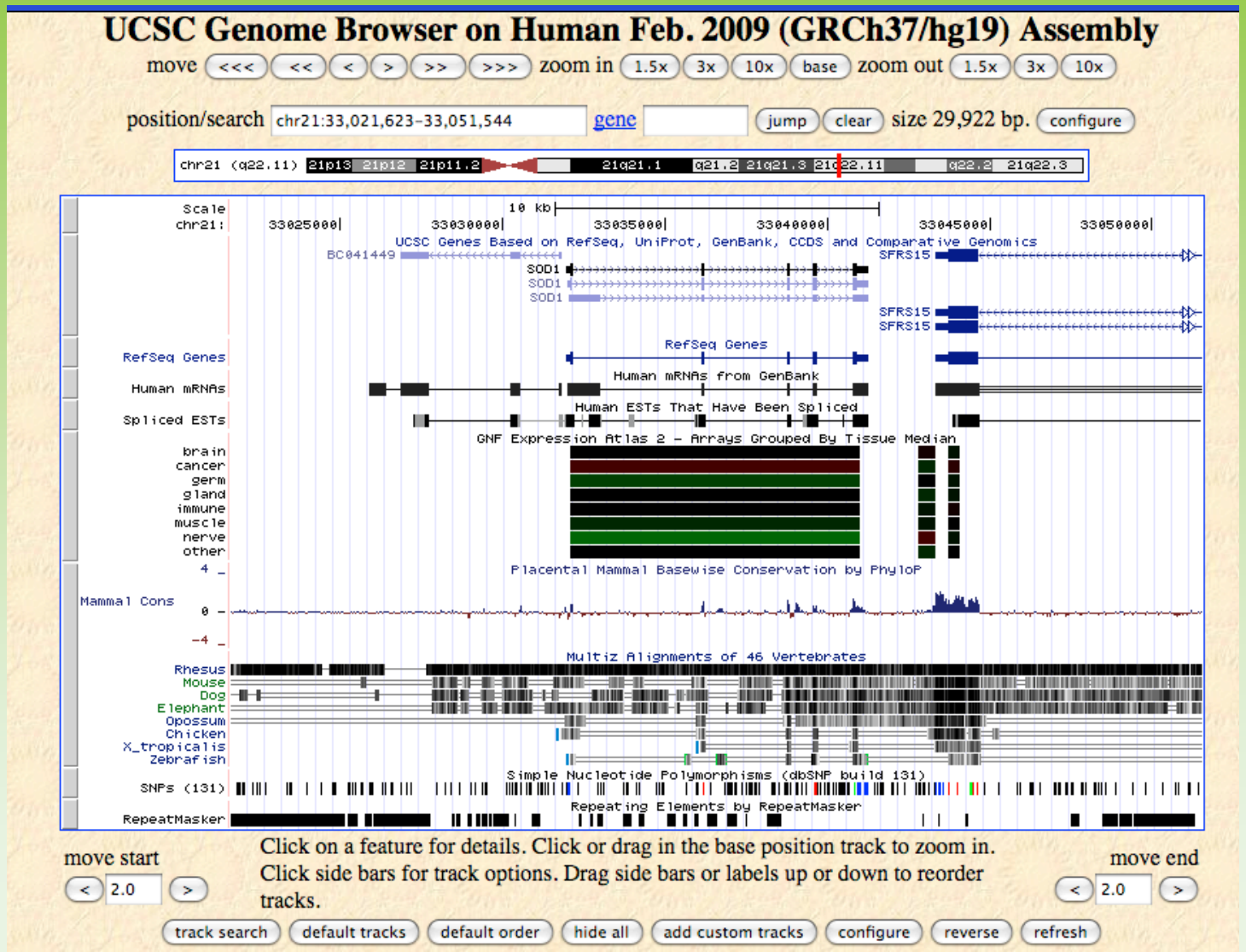
Visualization data formats

- Purpose-specific formats
  - SAM/BAM (read alignments)
  - BED, GFF, GFF3, GTF (features)
  - BedGraph, WIG (dense, continuous-valued data)
  - IGV (interval-based numeric data)
  - VCF (variant calls)

Visualization data formats

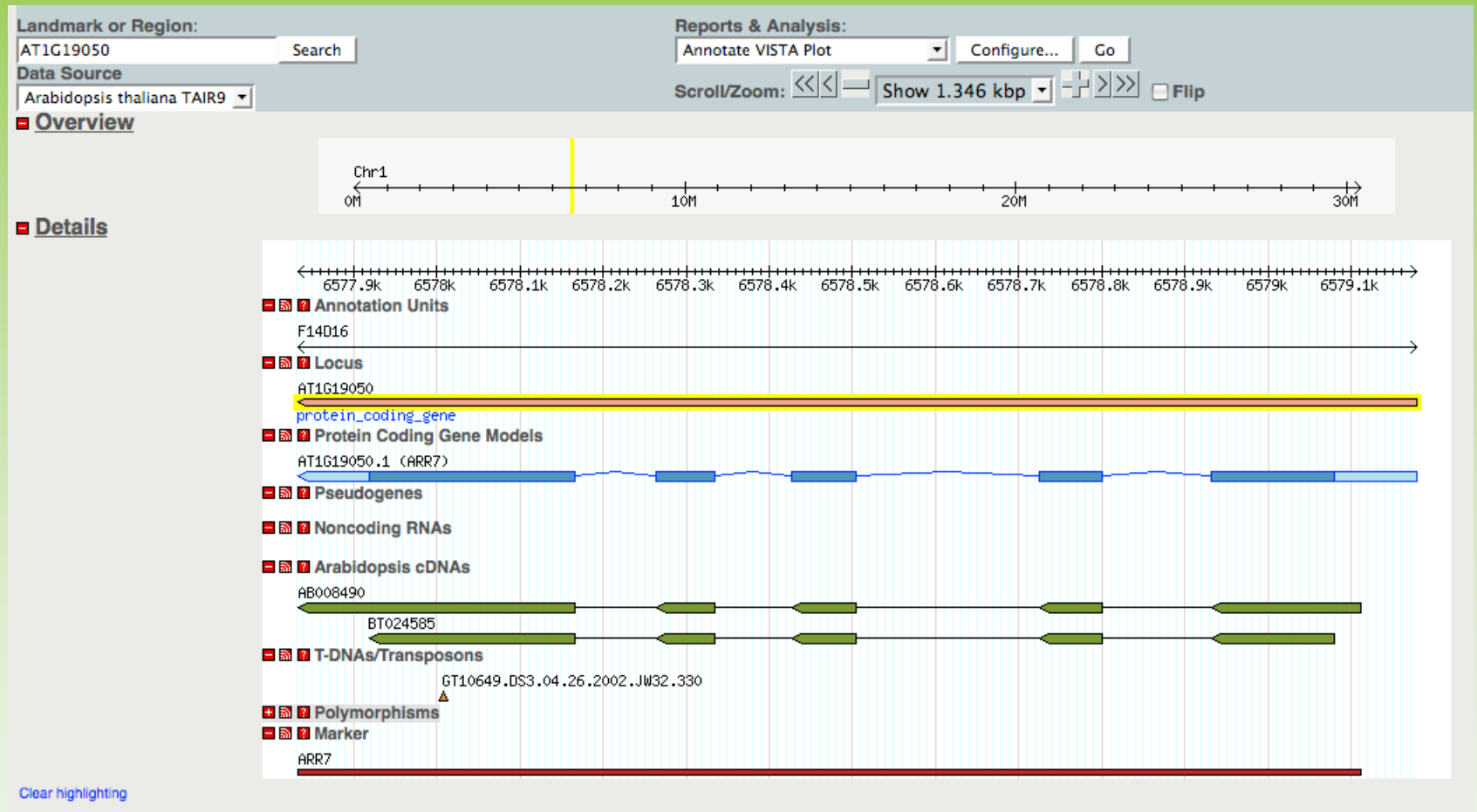
- Watch out for “0” or “1” based formats
  - Zero-based index: Setting start-end to 1-2 describes exactly one base, the second base in the sequence. (BED)
  - One-based index: Setting start-end to 1-2 describes two bases, the first and second in the sequence. (GFF, GTF, GFF3)
  - Refer to http://www.broadinstitute.org/igv/FileFormats

UCSC Genome Browser



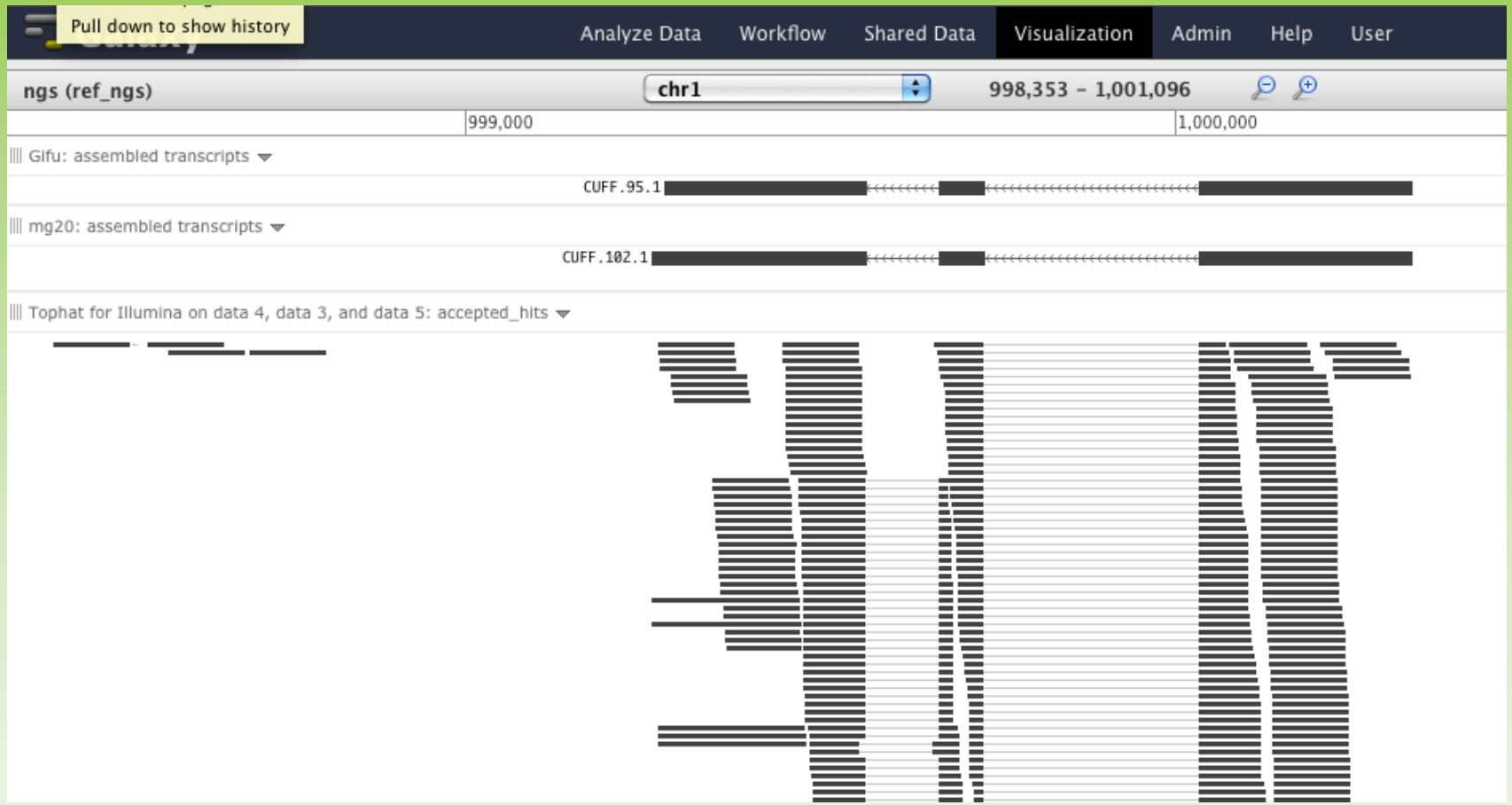
Visualization software

Arabidopsis GBrowse



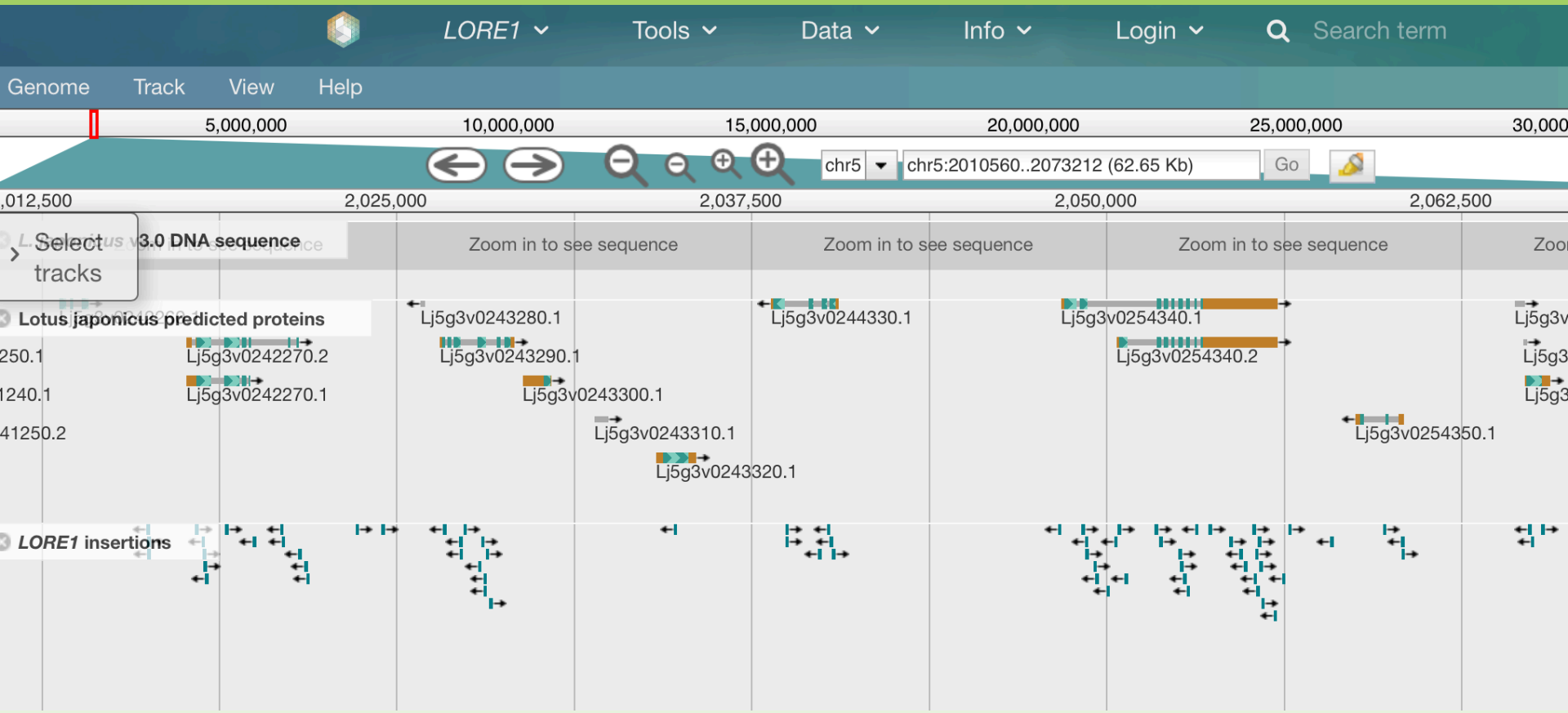
Visualization software

Galaxy



Visualization software

Lotus japonicus JBrowse



Visualization software

IGV

Visualization software

Simple quantitative data visualization

- Histograms for distributions
- Scatterplots for correlations