

The Generic Model Organism Database project

Scott Cain

scott@scottcain.net

GMOD project coordinator (among other things)

Ontario Institute for Cancer Research

October 19, 2022



About me

- PhD in (Bio)Chemical Engineering in the late 90s
- Decided at the end to call myself a bioinformatician
- Taught myself Perl (and C, and Java, and JavaScript, and Python)
- Worked at a few bio-related startups
- For 20 years, worked for Lincoln Stein (at CSHL then OICR) on
 - GMOD (Coordinator)
 - WormBase (Developer)
 - Alliance of Genome Resources (NIH funded MODs) (Group leader)
 - Project manager for a few bio-portal projects (Cancer and Covid)



Intro to GMOD

- Started in 2001, funded by the NIH to reduce redundant software development efforts at NIH-funded MODs (Rat Genome Database, Mouse Genome Informatics, WormBase, FlyBase, Saccharomyces Genome Database + plus Gramene (rice) and TAIR (arabidopsis) later)
- Funded 2 developers at each MOD to work together as a group to identify projects that could be made generic.
- Also funded a few meetings a year which resulted in building a community that continues to this day.



GMOD projects that have stood the test of time

- GBrowse (widely used genome browser, end of life a few years ago)
- CMap (genomic comparative mapping tool—limped along for a long time after the developer left, and some ideas have come along to replace it)
- JBrowse (a JavaScript genome browser, even more widely used)
- Apollo (initially a Java application for graphical feature editing, then a web-based tool based on JBrowse)
- Intermine (a data querying tool)
- Chado (a database schema)
- Tripal (a MOD web service built on Drupal)
- MAKER (a gene prediction tool)
- Galaxy (a web based computation pipelining tool)



...and some that didn't

- PubSearch (a reference tracking system)
- FlashGViewer (a flash-based genome browser)
- A “generic gene page” project (though Alliance of Genome Resources has kind of moved in this direction)

But the point is, we tried a lot of things, and a fair number of those were quite successful.



Types of projects we typically do

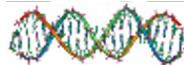
- Annotation creation/curation (both automated and manual)
- Data storage (databases and “queriers”)
- Visualization (browsers, whole websites)

What we don't do:

- Algorithms! (ie, no gene predictors)

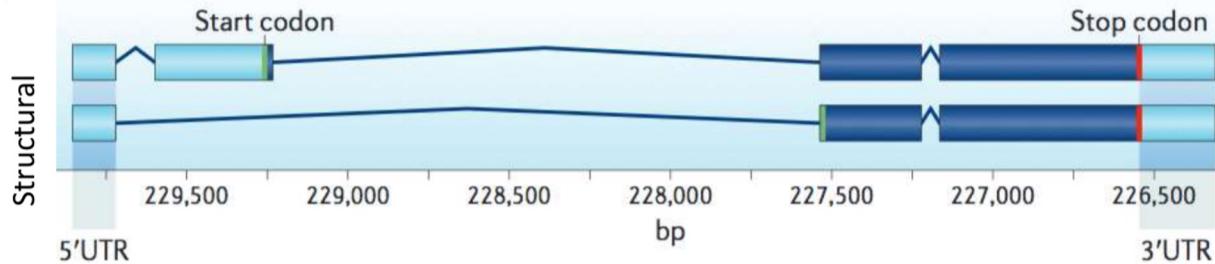


Genome Project Overview



Genome Project Overview

What is an Annotation?



Functional

Function	cAMP-dependent and sulfonylurea-sensitive anion transporter. Key gatekeeper influencing intracellular cholesterol transport.
Subcellular location	Membrane; Multi-pass membrane protein Ref.13 Ref.14.
Domain	Multifunctional polypeptide with two homologous halves, each containing a hydrophobic membrane-anchoring domain and an ATP binding cassette (ABC) domain.

D

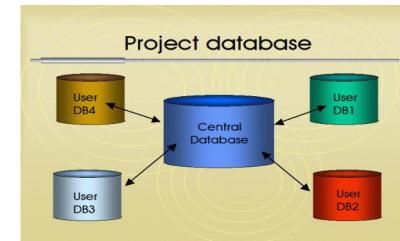
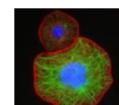
Genome Project Overview



SUCCESS



```
>Smg5  
MEVTFSSGGSSNASSECAIDGTTNRCRGL  
EPNNGTCTILSQEVKDLYRLSYTAKQLDD  
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
```



D

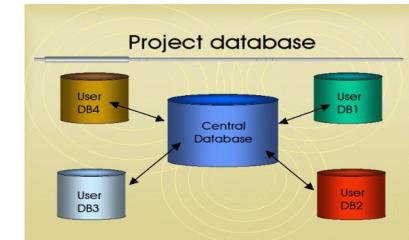
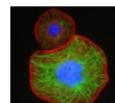
Genome Project Overview



Incorrect annotations poison every experiment
that uses them!!



```
>Smg5  
MEVTFSSGGSSNASSECAIDGGTNRCGL  
EPNNGTCTILSQEVKDLYRSLYTAKQLDD  
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
```



D



MAKER

an annotation pipeline and genome-database management
tool for second-generation genome projects

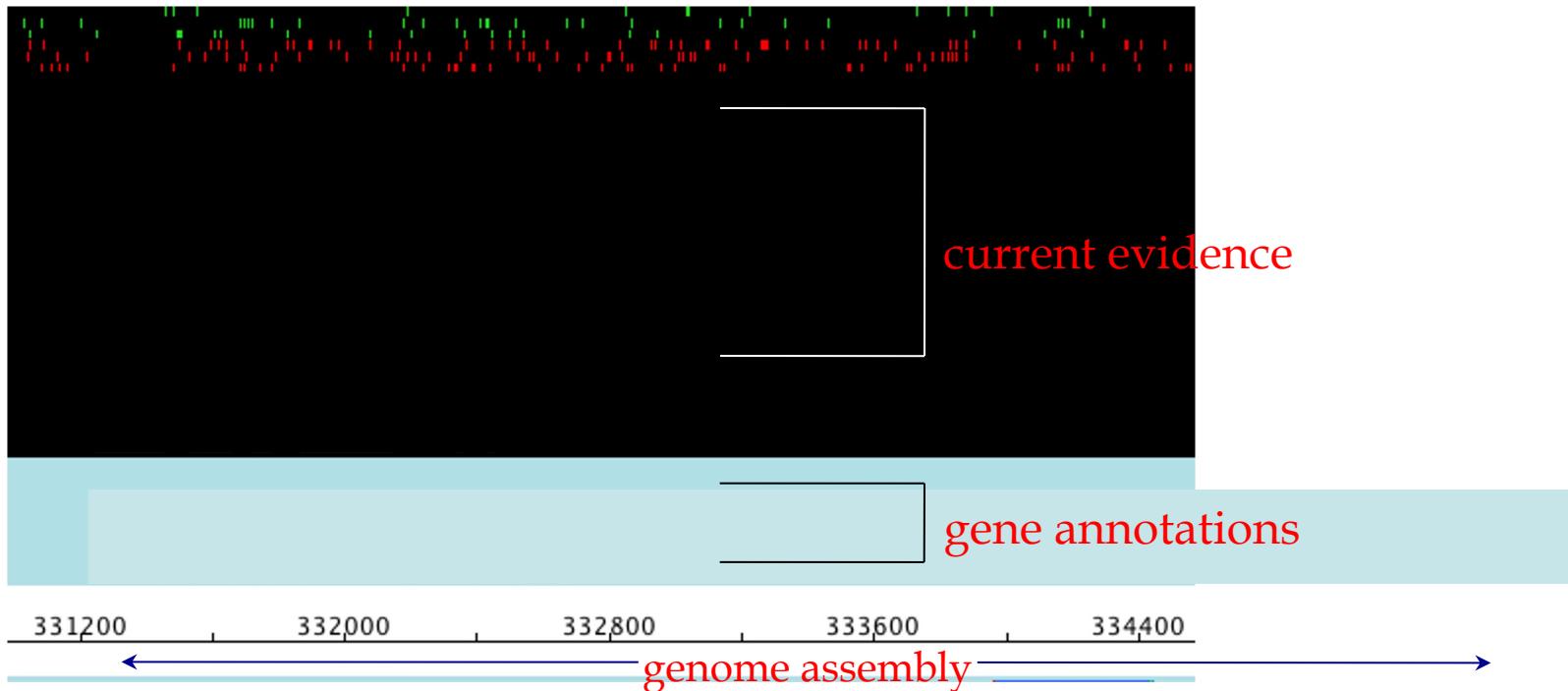


Easy-to-use by design

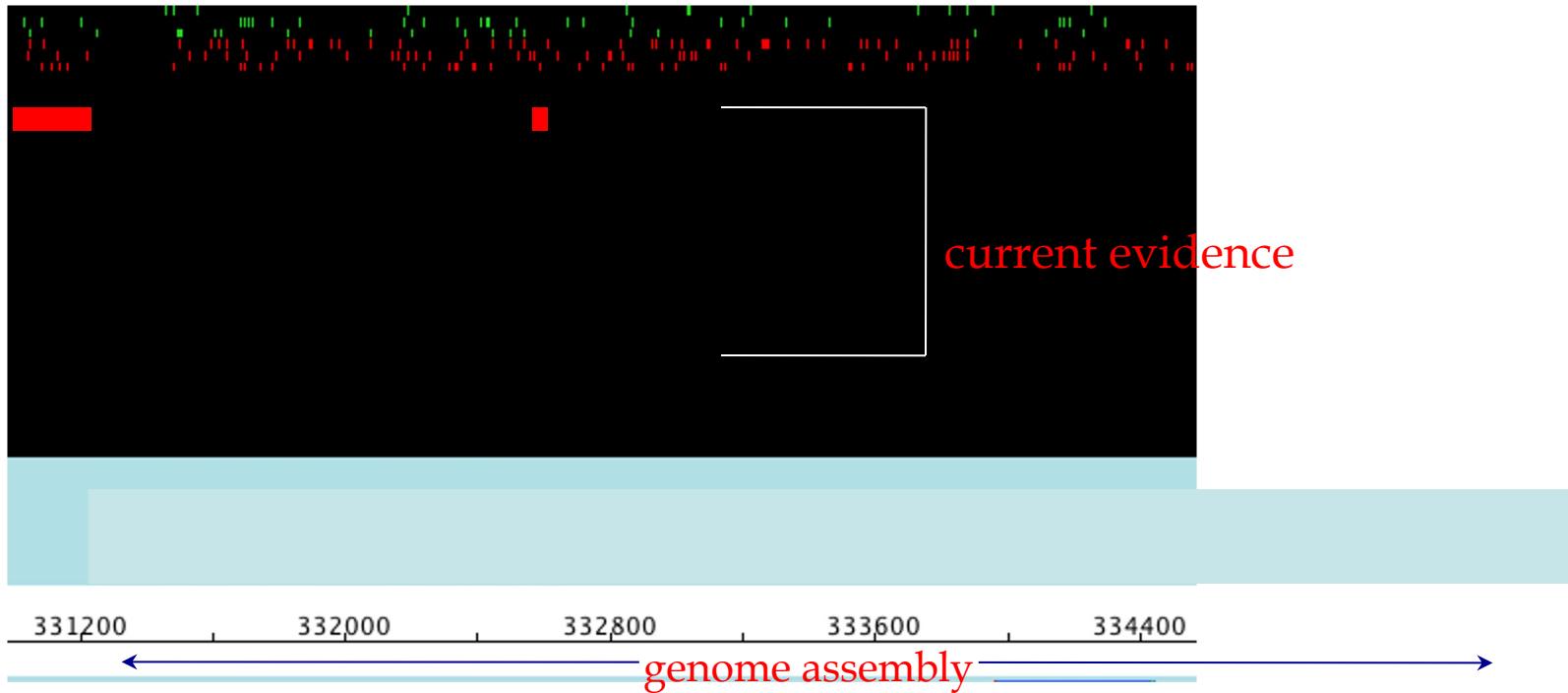
User Requirements:	Can be run by a single individual with little bioinformatics experience
System Requirements:	Can run on laptop or desktop computers (running Linux or Mac OS X)
Program Output:	Output is compatible with popular annotation tools like Apollo, GBrowse, and JBrowse
Availability:	Free open source application (for the academic community)



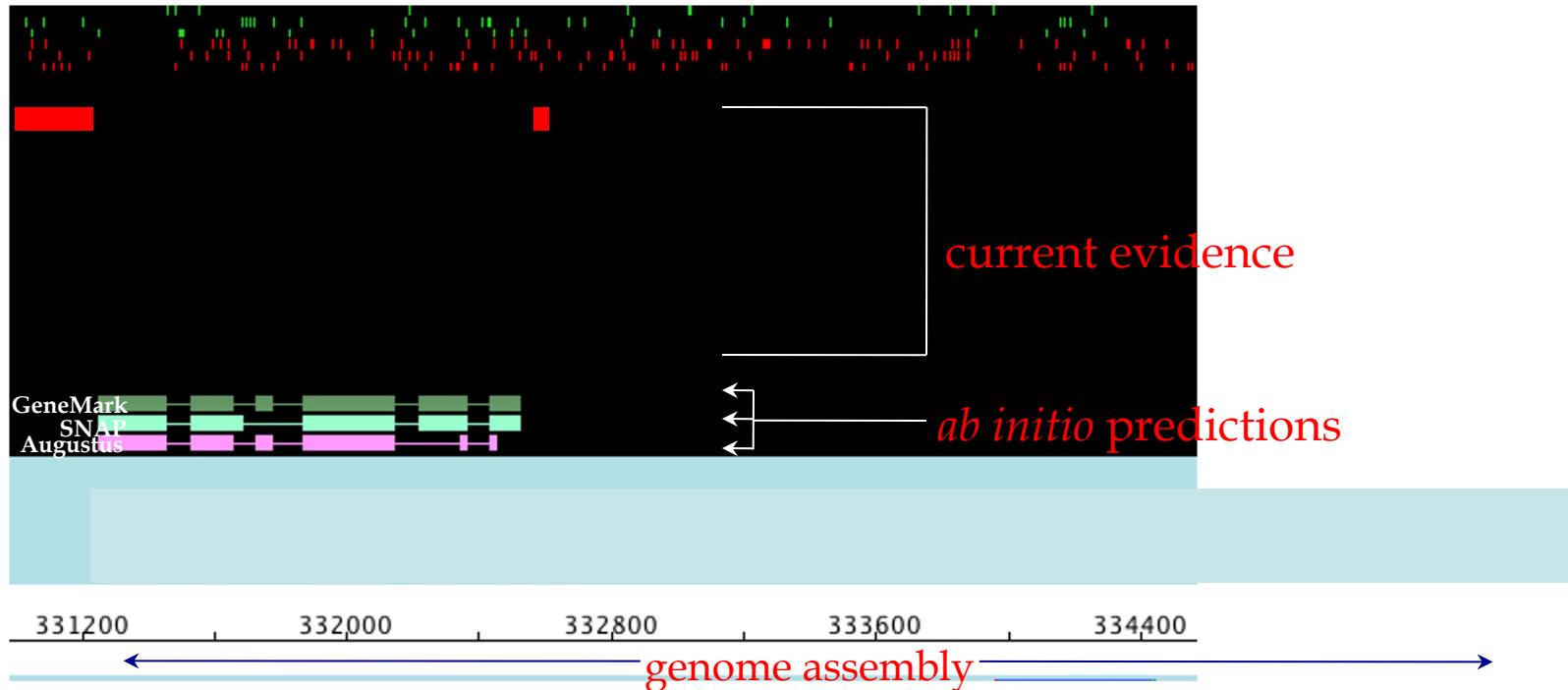
Annotating the Genome – Apollo View



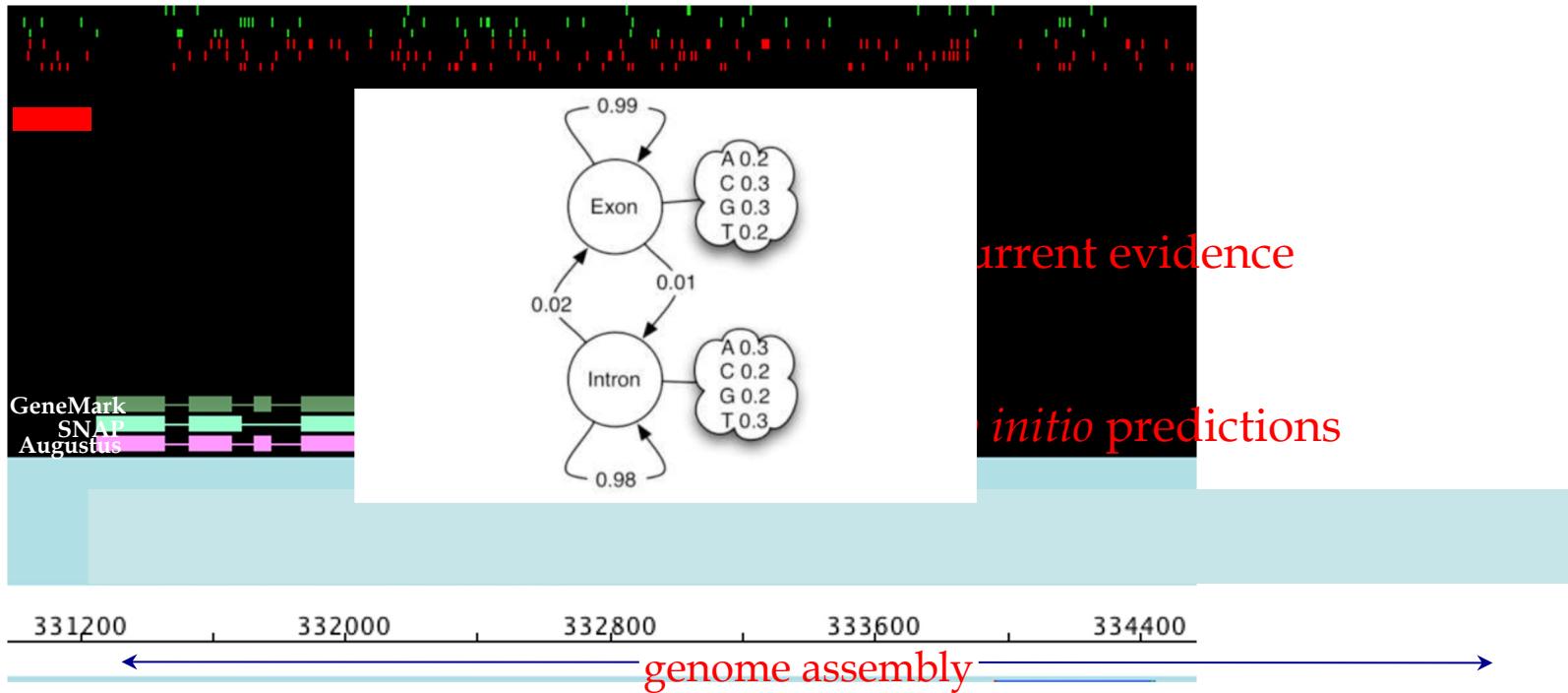
Identify and mask repetitive elements



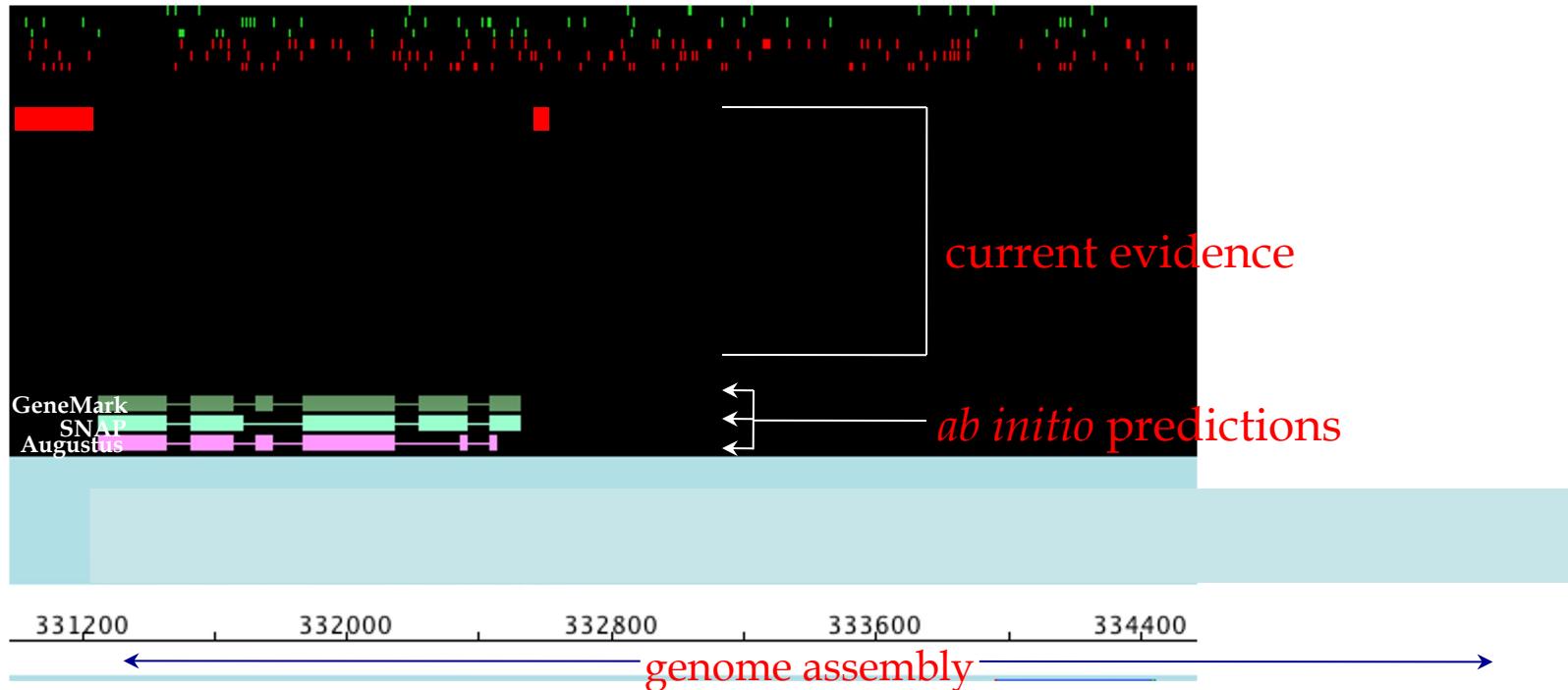
Generate *ab initio* gene predictions



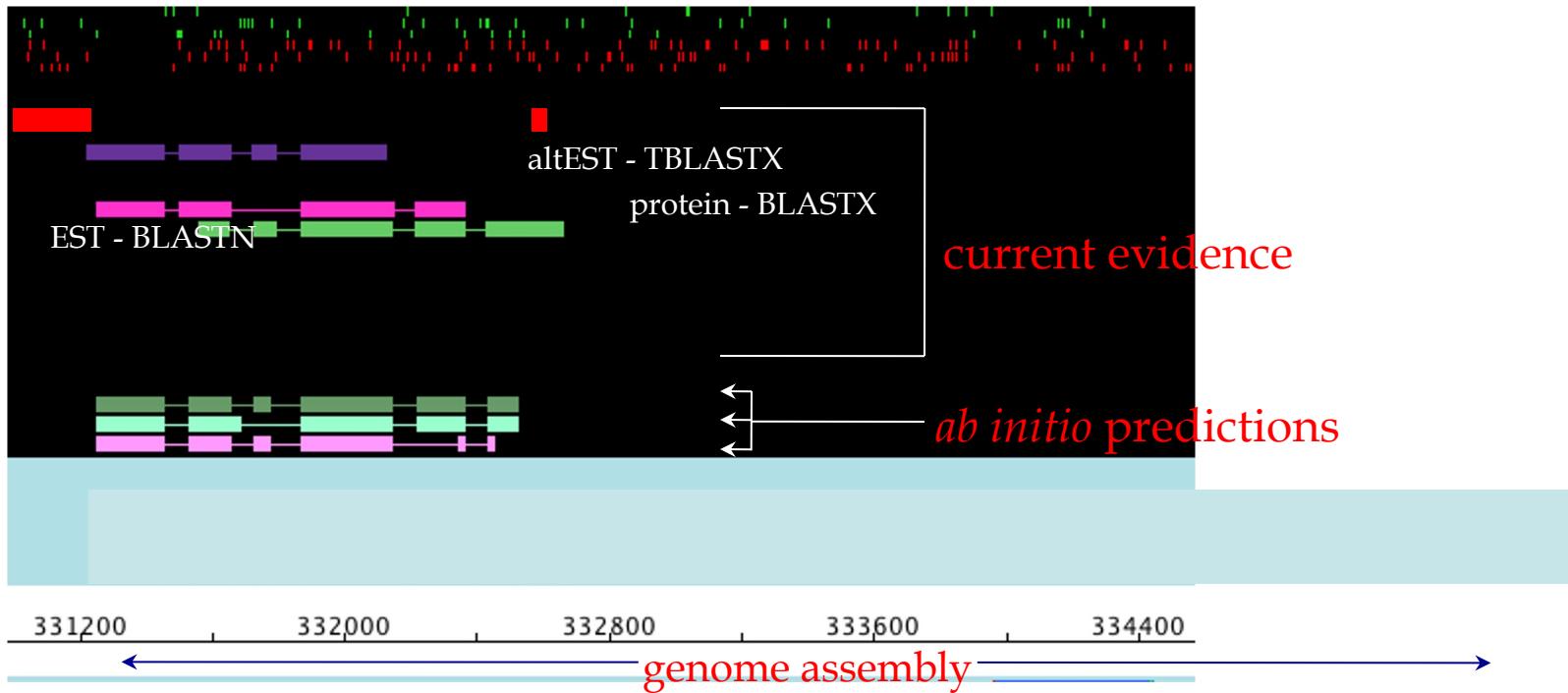
Generate *ab initio* gene predictions



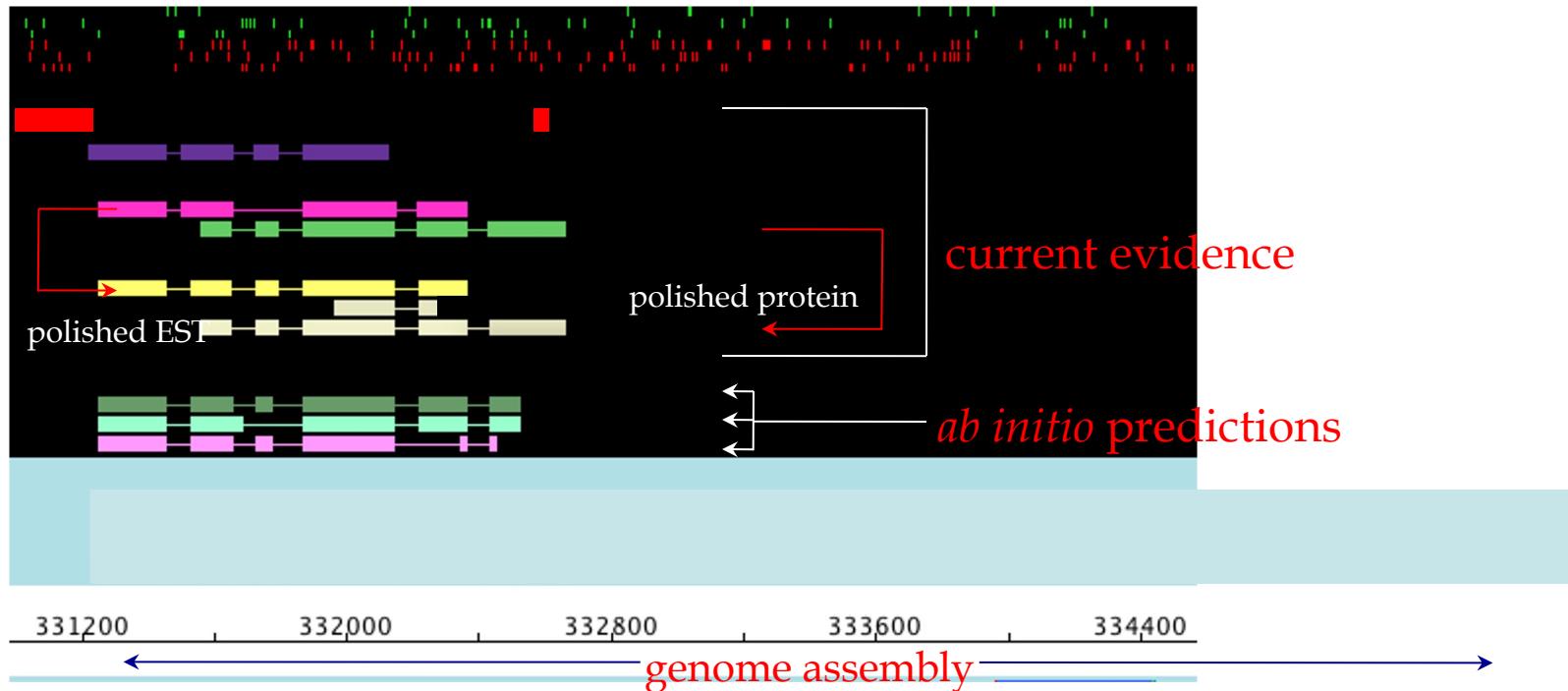
Generate *ab initio* gene predictions



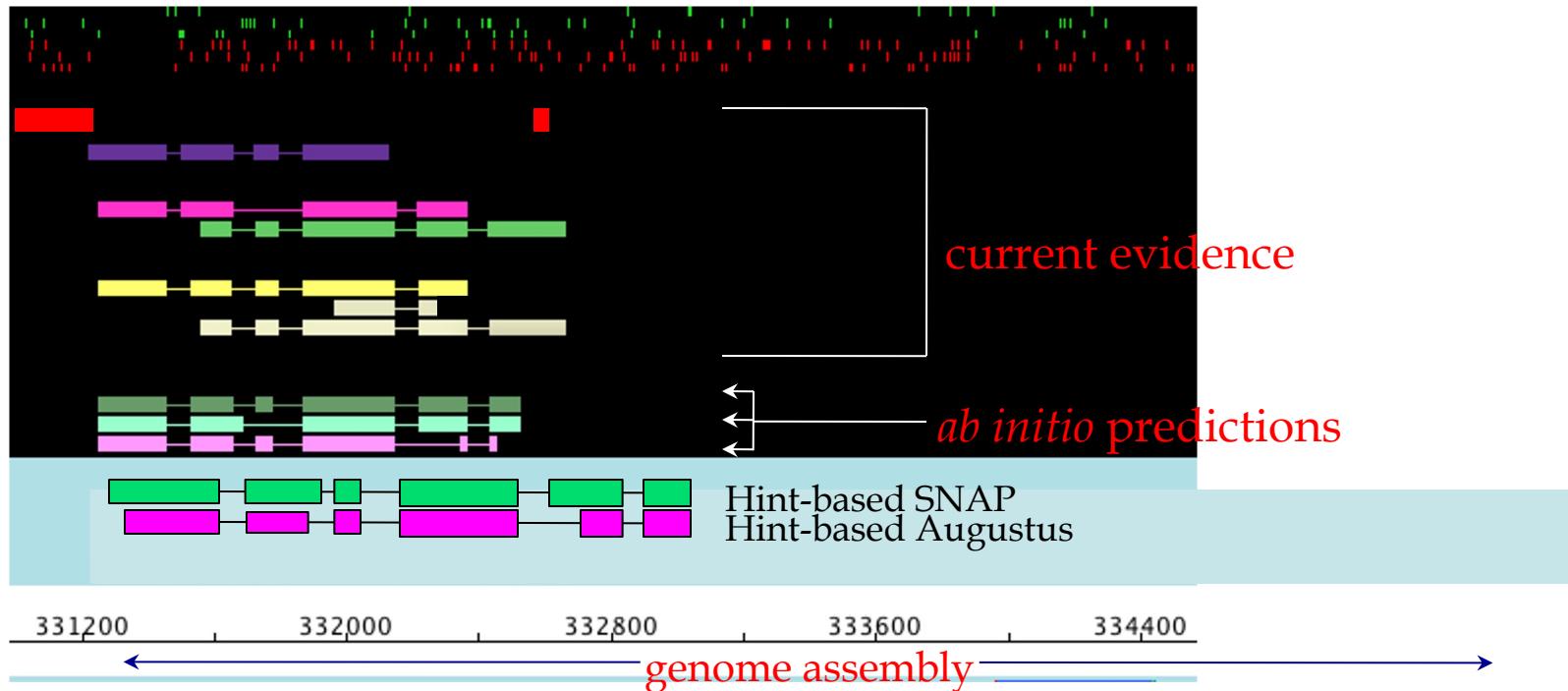
Align EST and protein evidence



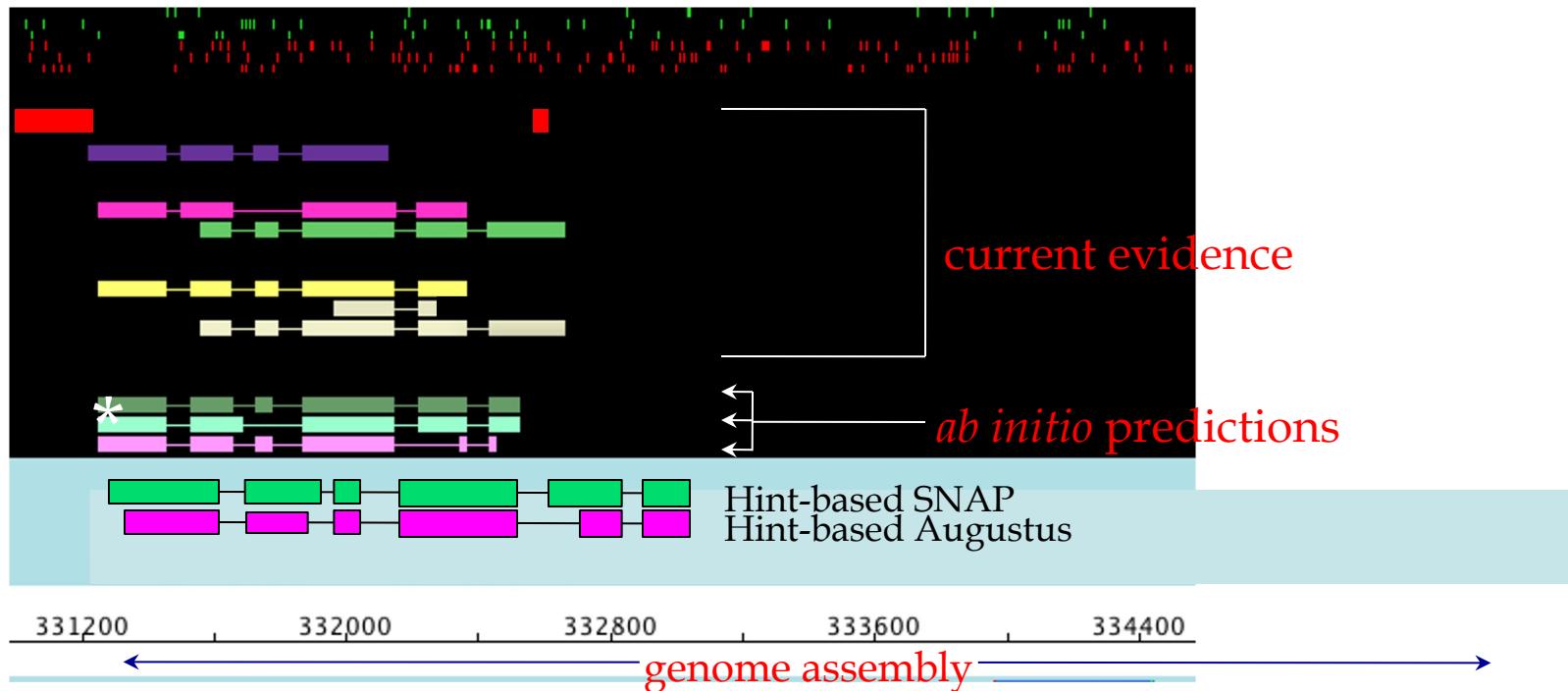
Polish BLAST alignments with Exonerate



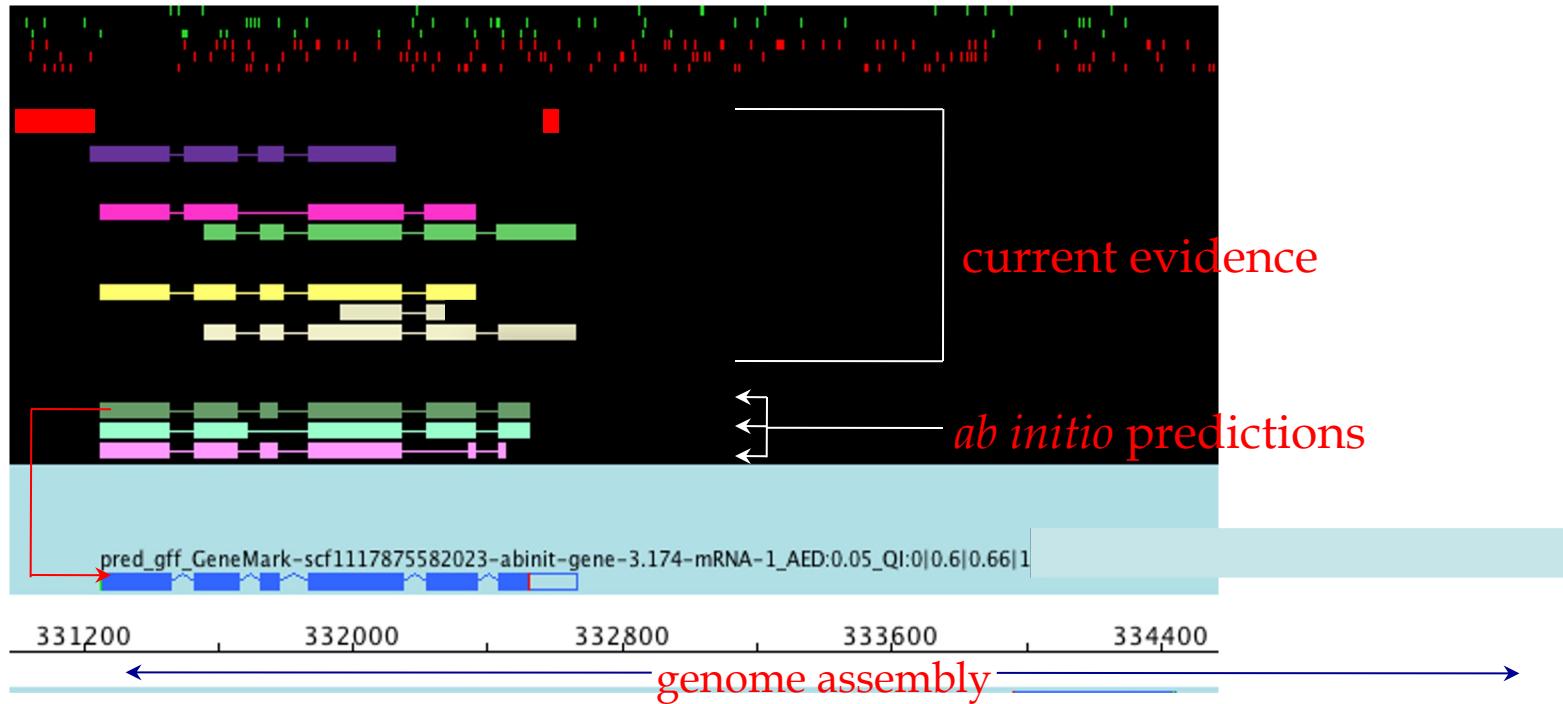
Pass gene-finders evidence-based ‘hints’



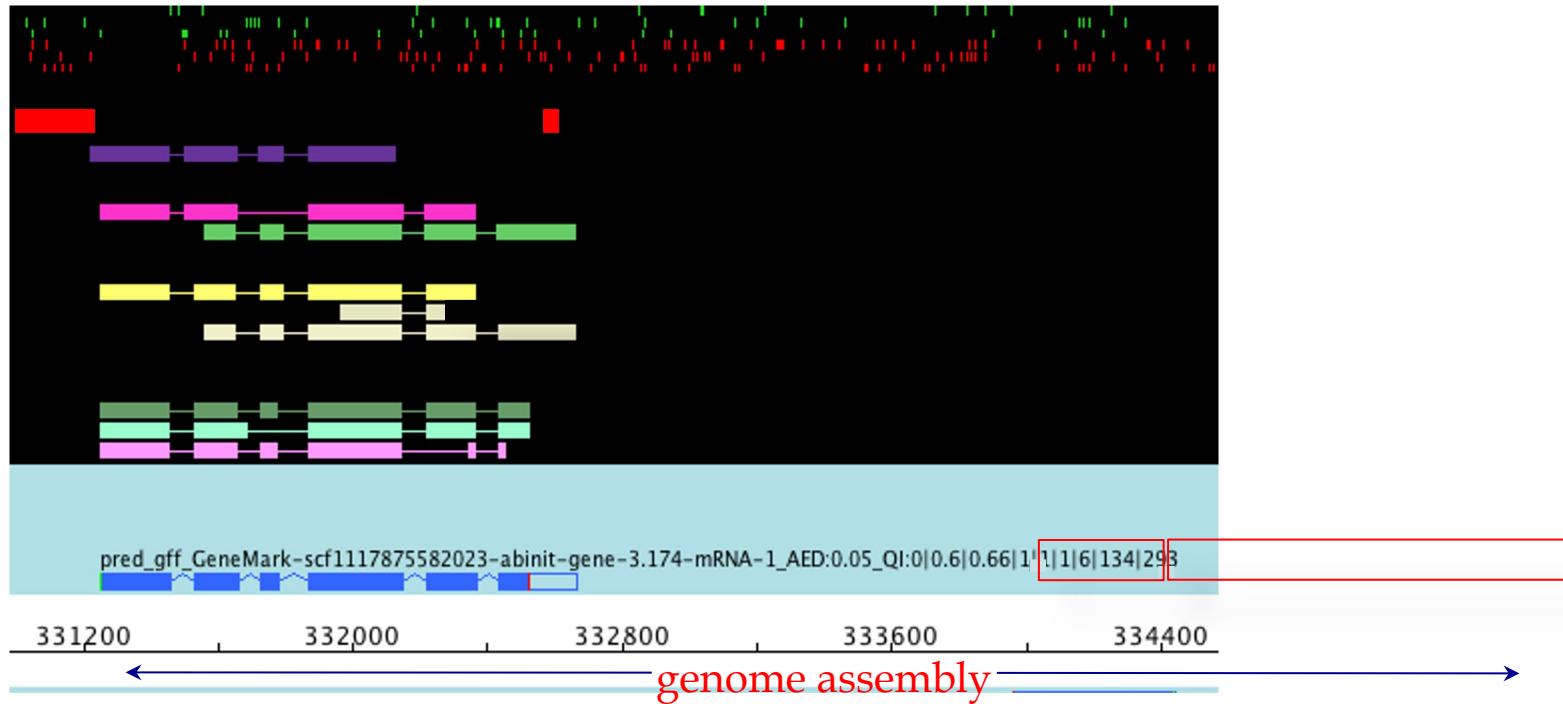
Identify gene model most consistent with evidence



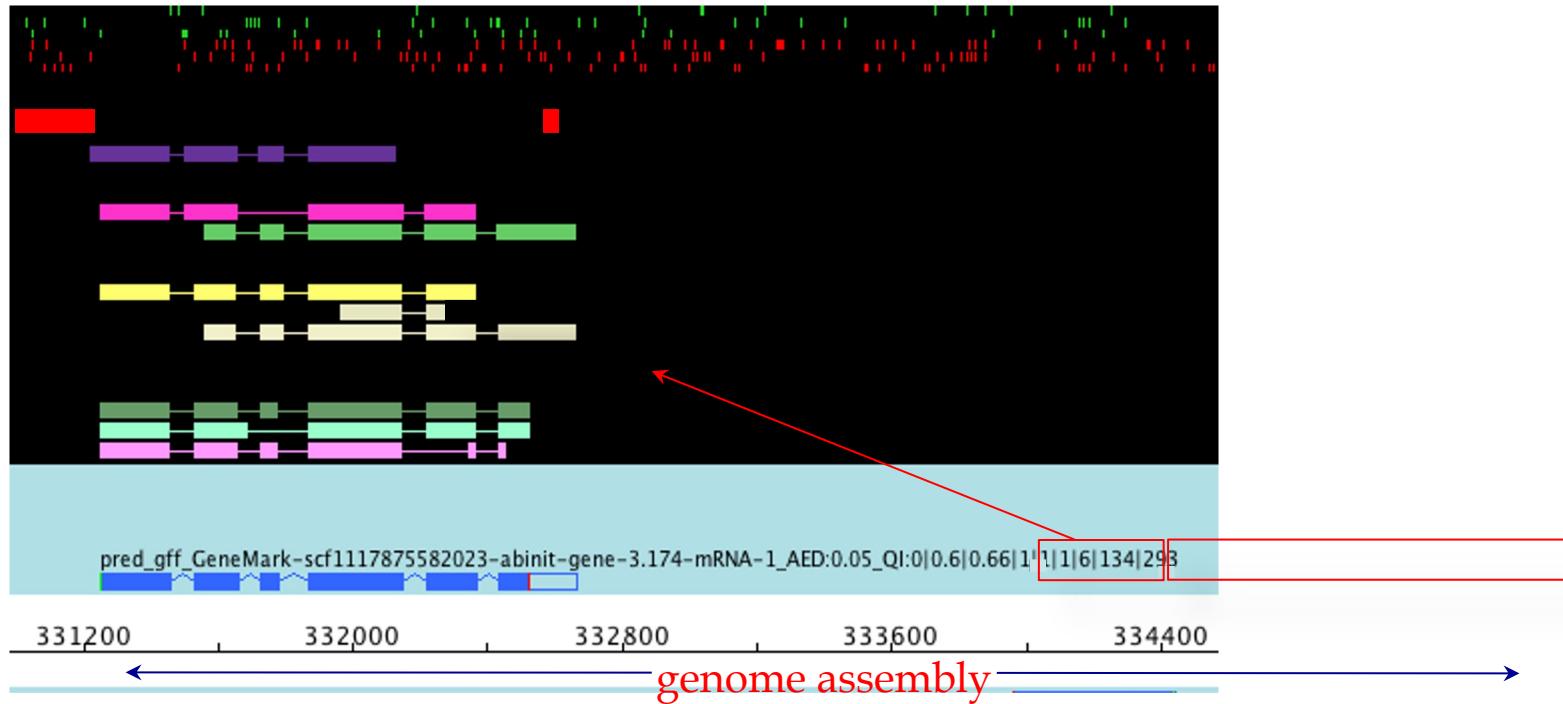
Revise it further if necessary; create new annotation



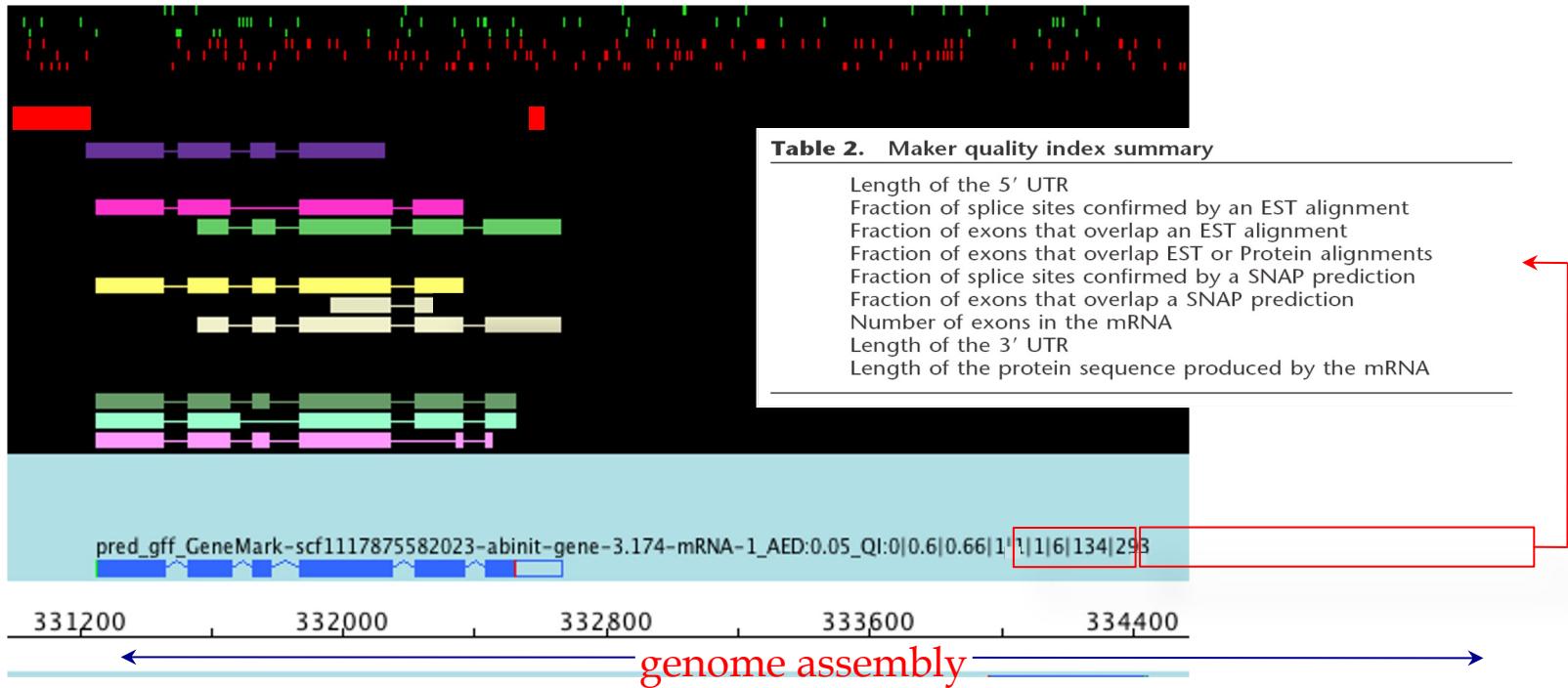
Compute support for each portion of gene model



Compute support for each portion of gene model



Compute support for each portion of gene model



bin — bmoore@derringer:/data1/genomes/Homo/sapiens/36.2 — ssh

```
scaffold00080 repeatmasker match_part 723561 723736 914 + . ID=scaffold00080:hsp:4987;Par
scaffold00080 repeatmasker match 724950 725171 495 + . ID=scaffold00080:hit:3405;Name=specie
scaffold00080 repeatmasker match_part 724950 725171 495 + . ID=scaffold00080:hsp:4988;Par
scaffold00080 repeatmasker match 726925 727293 724 + . ID=scaffold00080:hit:3406;Name=specie
scaffold00080 repeatmasker match_part 726925 727293 724 + . ID=scaffold00080:hsp:4989;Par
scaffold00080 maker gene 56197 58302 . + . ID=ACEP_00015614;Name=ACEP_00015614;Alias=mak
scaffold00080 maker mRNA 56197 58302 . + . ID=ACEP_00015614-RA;Parent=ACEP_00015614;Name
scaffold00080 maker exon 56197 56274 . + . ID=ACEP_00015614-RA:exon:126;Parent=ACEP_0001
scaffold00080 maker exon 56569 56584 . + . ID=ACEP_00015614-RA:exon:127;Parent=ACEP_0001
scaffold00080 maker exon 56797 56906 . + . ID=ACEP_00015614-RA:exon:128;Parent=ACEP_0001
scaffold00080 maker exon 57851 57941 . + . ID=ACEP_00015614-RA:exon:129;Parent=ACEP_0001
scaffold00080 maker exon 58067 58302 18.726 + . ID=ACEP_00015614-RA:exon:130;Parent=ACEP_0001
scaffold00080 maker CDS 56197 56274 . + 0 ID=ACEP_00015614-RA:cds:124;Parent=ACEP_00015
scaffold00080 maker CDS 56569 56584 . + 0 ID=ACEP_00015614-RA:cds:125;Parent=ACEP_00015
scaffold00080 maker CDS 56797 56906 . + 2 ID=ACEP_00015614-RA:cds:126;Parent=ACEP_00015
scaffold00080 maker CDS 57851 57941 . + 1 ID=ACEP_00015614-RA:cds:127;Parent=ACEP_00015
scaffold00080 maker CDS 58067 58302 . + 2 ID=ACEP_00015614-RA:cds:128;Parent=ACEP_00015
scaffold00080 maker gene 2797 3250 . - . ID=ACEP_00015615;Name=ACEP_00015615;Alias=mak
```

GFF3

bin — bmoore@derringer:/data1/genomes/Homo/sapiens/36.2 — ssh

```
>ACEP_00015614-RA protein AED:0.401129943502825 QI:0|0|0|0.2|1|1|5|0|177
MEKSDDYQDHVYIQFLEVYMMYFVHEKQYLYKLVLWKRITLNSKYPPIGVIAFFPIVF
IRHPDDLKTILSNPKHKKSSFYDNIKPWLGTSLTSEDCTLSHFIPMLSNRHFTIVHCI
VHVLGTKWQLQRKILISTFHFDILNQFVEIFEKESENKMIKSLKNAEGTVVKDLSSFI
>ACEP_00015615-RA protein AED:0.223684210526316 QI:60|0|0|0.5|0|0|2|0|193
MDTQQKHREIPTEPGEINSVISFQCISCCDLSYCNIESTPTNATNAIYSRQRRAKSJK
RKRPGRNGVDAATRLYGSSLWLPAASLRYAYHC
>ACEP_00015613-RA protein AED:0.191135958515638 QI:9|0|4|0.33|0.83|0.2|0.5|6|0|238
MEKSDNSLTYQDHYSIFGSVYVLFGSREILWKIITLNSKFLRTFYPIGIIRAFFPIVSI
RHPDDLKSFSFYDNIKPWLGTSLINEGTWKQLQRKILIPTFHDILNQFVEIFEKESEN
IKSLKNAEGTVVKDLSSFISEYTLNAICESYLIPIAGTVLHNINGVHTDPNEWRYFQIQR
YLILIDFCPRGEIVTLLTHIYHSVLVDHVIVSNLRFSKFDHERSGIVRE
>ACEP_00015612-RA protein AED:0.469714351691491 QI:10|0|0|0.3|0.7|0.2|0.5|6|0|249
MYFVNPNFYANASVTNHRNGRTSLTELSSKVKREQLRNILTARR
RILSSRRPRPGDIISKDPQDVLETQETRNGIYIANARTEISIKEEIPQ
INLNATGEELESRTTKTSKNESLSPTEKIEINISSLVSTPATNQKLGAEMRKKNIVSMP
ESLYNHFRPVESNIPVEDMSQFLYFGQKLQPDALNVTSNNSVETTSTNPNTSSRRRYST
KRFTATIATPMEEIMNEEMNIALETLERRTVEKNQVSRIKNTFRSSGNGLYYRKPRPTAD
VVSNIIPGNESSPGIIGGGSETKSEMNNTVDDKFHRSDSAHADGSRDRVPLERKNAAHVSE
VSMTTQIPRQSDVEARVIVESEIGAKLSNKTTIASNGIVESLQRINDFGKDNEKKKITEA
EAGNVEKSEHKLMSVERTISTSQRESLTRESRFDSTETMSLVPSRAESSTLRLIVVPDNSFG
LSCGATECERGCHNTYK'CGCVV'DPTABPDRBPCGCCAACCGCAATTGCCCCGPPGCGCGGFG
```

FASTA



MAKER Resources

<http://yandell-lab.org/software/maker.html> Docs and demo





Galaxy

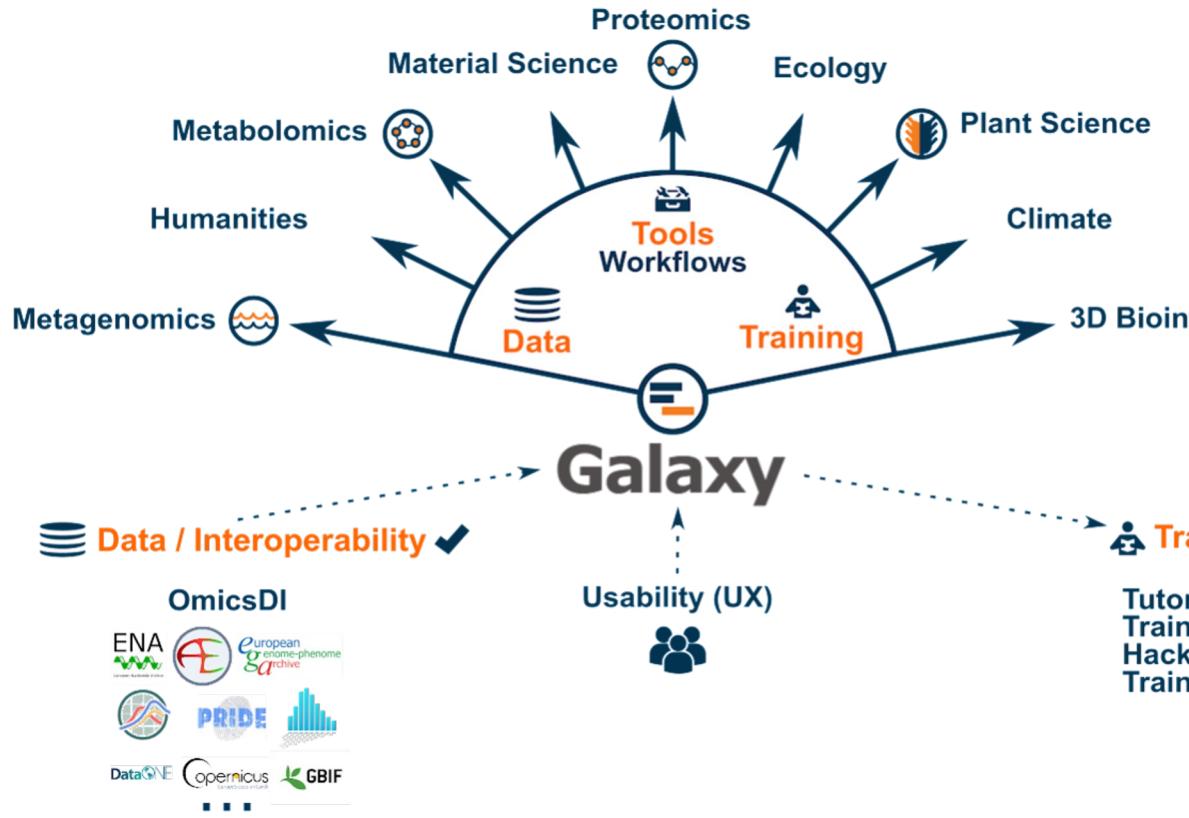
PROJECT

Galaxy is an **open**, web-based platform for
accessible, reproducible, and **transparent**
computational research

<https://galaxyproject.org>



A Data Analysis Gateway for Everyone!



Web-based User Interface (+ API access)

Galaxy

Analyze Data Workflow Visualize Shared Data Admin Help User Using 4.1 TB

Certain large memory tools are temporarily running with reduced memory (RNA STAR, SPades, Unicycler) or have been temporarily disabled (Trinity).

Tools

- search tools
- Inputs
- Get Data
- Collection Operations
- Expression Tools
- GENERAL TEXT TOOLS
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Datamash
- GENOMIC FILE MANIPULATION
- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over
- COMMON GENOMICS TOOLS
- Operate on Genomic Intervals
- Fetch Sequences/Alignments
- GENOMICS ANALYSIS
- Assembly
- Annotation

COVID-19: PE Variation

Galaxy / Europe

Analyze Data Workflow Visualize Shared Data Help User Using 4.6 TB

Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below. If you would like to perform self-mapping select 'history' here, then choose your input file as reference.

Using reference genome

Human Dec. 2013 (GRCh38/hg38) (hg38)

Galaxy Galaxy @ Belgium

Analyze Data Workflow Visualize Shared Data Admin Help User Using 4.6 TB

History

- cover-19 original data
- 99: data 97 converted to fastqsanger (READ2)
- 98: data 97 converted to fastqsanger (READ1)
- 97: MergeSamFiles on data 96, data 95, and data 94
- 94: Merge BAM dataset
- 93: Filter SAM or BAM, output S and BAM on collection 89: bam
- 89: Map with BWA-MEM on collection 60 (mapped reads in BAM format)
- 88: MultiQC on data 86, data 78, and data 70: Web page
- 87: MultiQC on data 86, data 78, and data 70: Stats
- 62: fastp on collection 3: JSON Report
- 61: fastp on collection 3: HTML Report

SRR10902284

88.1 MB

format: fastqsanger.gz, database: ?

SRR10948474

285.4 MB

format: fastqsanger.gz, database: ?

SRR10948550

142.4 MB

format: fastqsanger.gz, database: ?

Tools



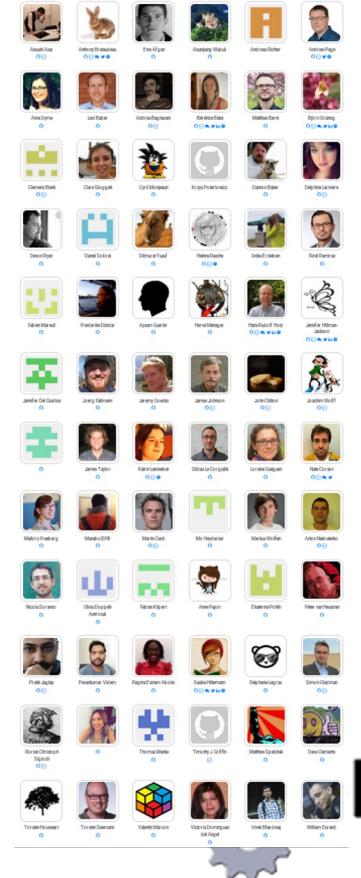
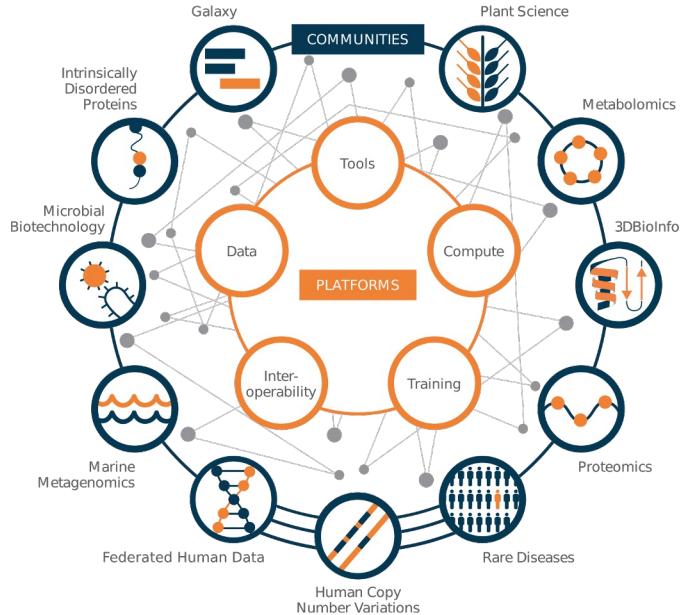
- ~8000 tools available
- App Store-like distribution system
- Can run in Conda, Containers, modules, etc.
- DOTADIW philosophy

BIOCONDA®



Communities sharing one coherent framework

- rna.usegalaxy.eu
- clipseq.usegalaxy.eu
- metagenomics.usegalaxy.eu
- hicexplorer.usegalaxy.eu
- cheminformatics.usegalaxy.eu
- proteomics.usegalaxy.eu
- imaging.usegalaxy.eu
- metabolomics.usegalaxy.eu
- ecology.usegalaxy.eu
- nanopore.usegalaxy.eu
- singlecellomics.usegalaxy.eu
- humancellatlas.usegalaxy.eu
- virology.usegalaxy.eu
- climate.usegalaxy.eu
- streetscience.usegalaxy.eu
- mi.usegalaxy.eu
- annotation.usegalaxy.eu



Galaxy resources

<http://galaxyproject.org/>

<http://usegalaxy.org/>

And especially

<http://usegalaxy.eu/>



InterMine info

- Takes a variety of biological data in, denormalizes it, and makes it queriable in a very fast, powerful way.
- Used by many MODs (WormMine, YeastMine, FlyMine, etc)
- <http://intermine.org>

WormBase WormMine WS286 Intermine data mining platform for *C. elegans* and related nematodes

Home Templates Lists QueryBuilder Regions Data Sources API MyMine Log in

Search WormMine e.g. aap-1, WP:CE18491 GO

Search
Search WormMine. Enter names, identifiers or keywords for genes, proteins, transcripts, ontology terms etc.
e.g. aap-1, WP:CE18491
SEARCH

Analyze a list
Enter a list of identifiers of the same type. (aka: all genes, or all proteins).
Gene
e.g. acr-10, unc-26, hlh-2, WBGene00002299, WBGene00004323, WBGene00002992
advanced
ANALYSE

First Time Here?
WormMine integrates many types of data for *C. elegans*. You can run flexible queries, export results and analyse lists of data.
TAKE A TOUR

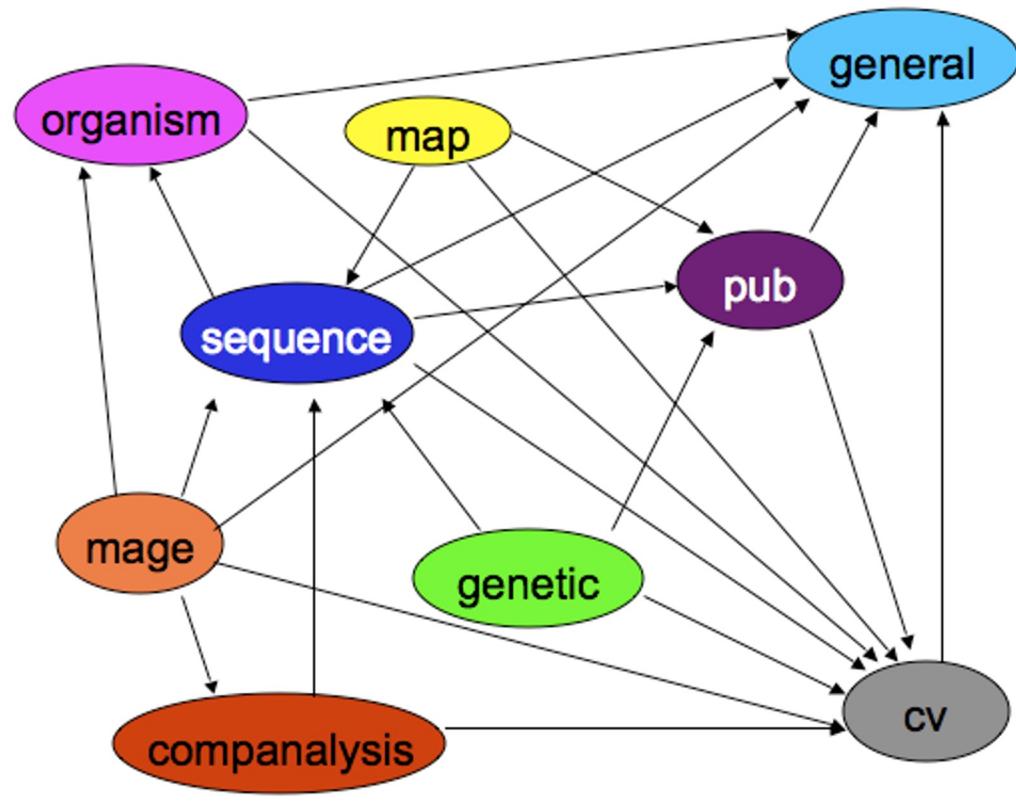


Intro to Chado

- Chado is a highly normalized relational database schema (most often used in PostgreSQL)
- Designed to support multiple organisms' genomes (organism agnostically)
- Makes heavy use of ontologies and controlled vocabularies to allow a wide variety of data and metadata types.
- Originally developed (and still in use by) FlyBase.



Chado diagram



Chado is most often used in conjunction with ...





What is Tripal?

An open-source Biological Database toolkit that

Aids in the creation of community-centric, data-driven websites



Facilitates FAIR data sharing

- Ontology-Focused
- Consistent web services



More features through shared development

- Focus on what makes your community unique

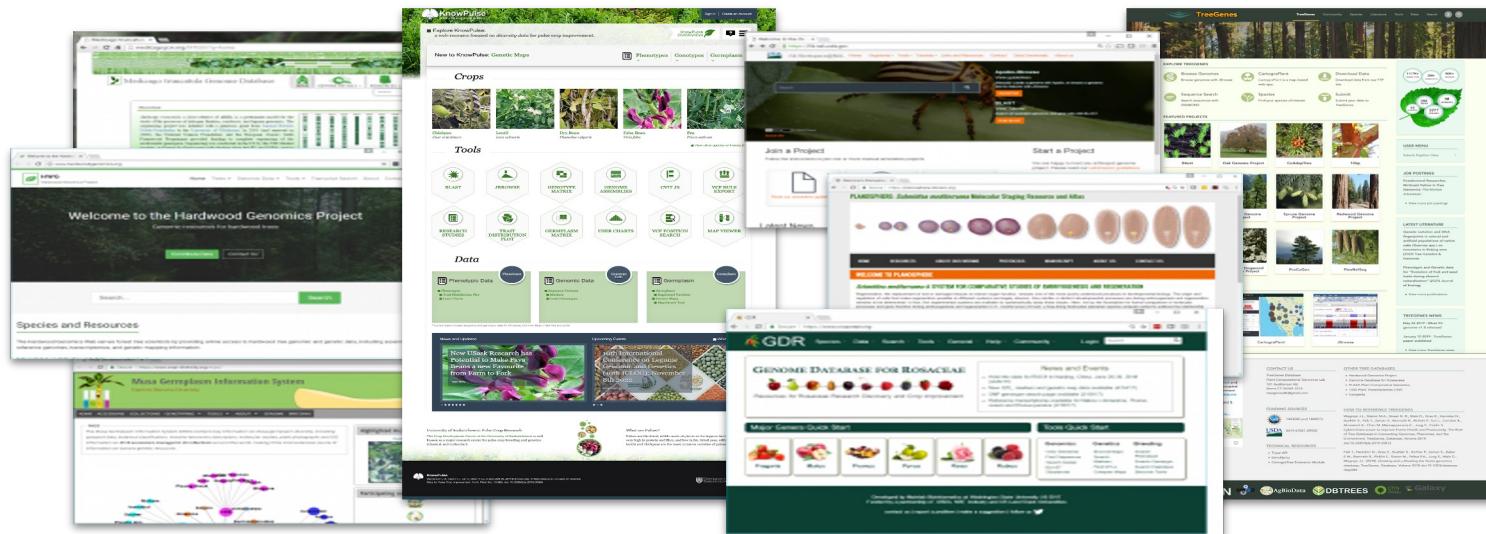


D



Tripal Community

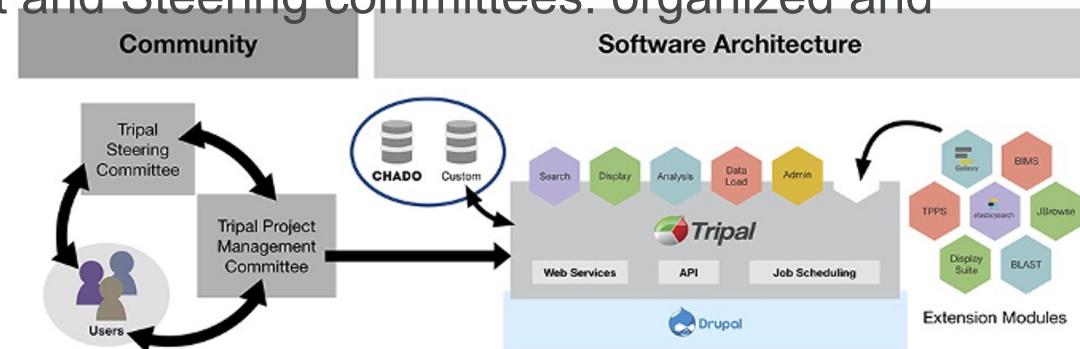
- Over 125 sites report using Tripal
- Species include plants, animals, insects, bacteria





Tripal Structure

- Easily extendable with rich APIs, documentation and core services
- Open-source to the core: integrates with and extends many other open source projects including Drupal, Chado, Galaxy, JBrowse, Elastic Search, etc.
- Community developed with users and developers forming the Project Management and Steering committees: organized and community powered





Tripal Features



Rich Pages for biological data.

- Extensive, flexible metadata
- Links to related data, data summaries, context
- Bulk Import or individual creation + curation



Flexible Searching makes it easy to use rich metadata to focus your results.



Robust security managed by the world-wide Drupal community.

- Keep data private until published,
- Share with collaborators using user roles + permissions



Both science and community focused extensions through Tripal and Drupal to customize your site.

D

Tripal Resources

<http://tripal.info/>

<http://tripal.info/join/slack>

Some Tripal examples:

<https://knowpulse.usask.ca>

<https://www.rosaceae.org>

<https://planosphere.stowers.org/smedgd>

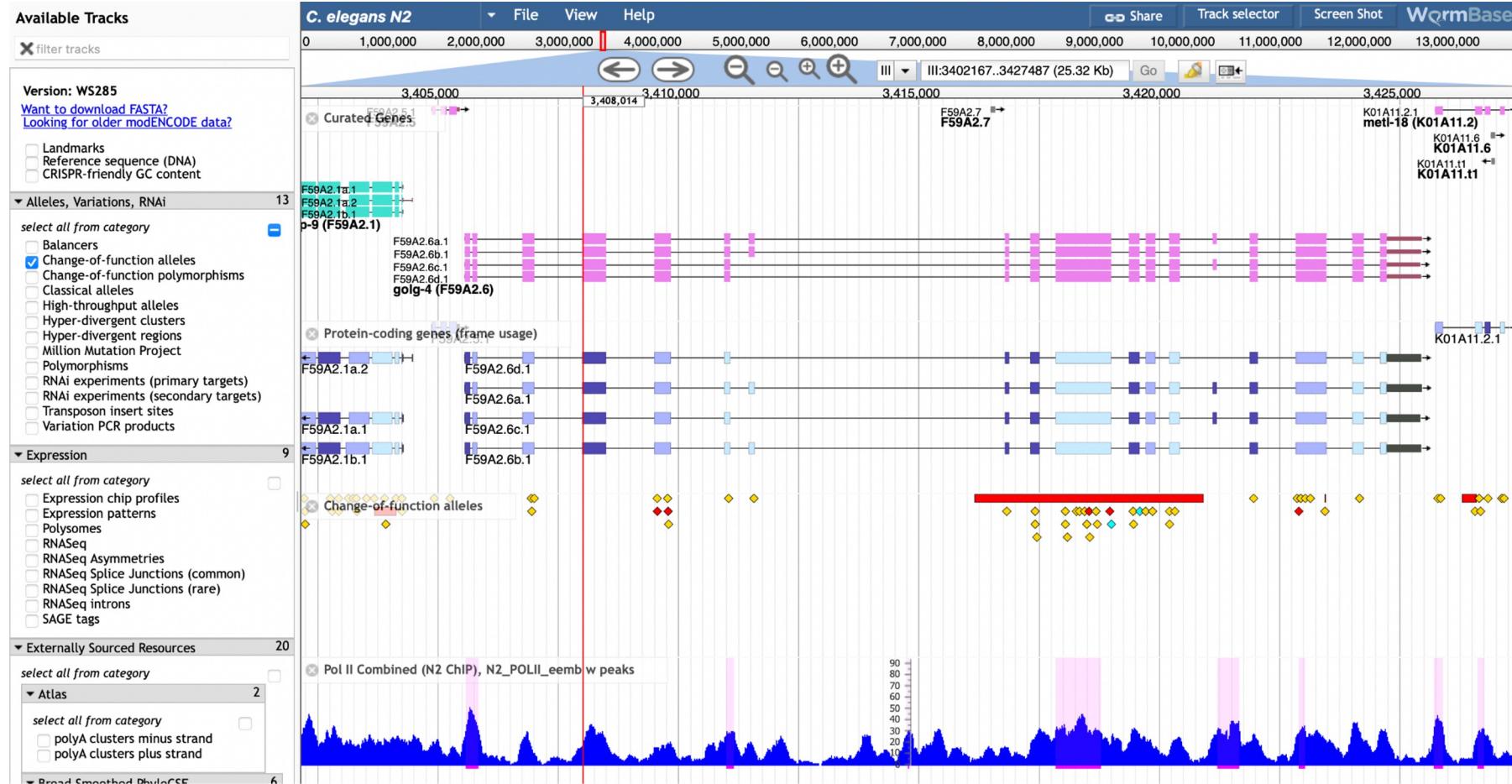


Intro to JBrowse

- A JavaScript-powered genome browser (think IGV or UCSC genome browser, but faster, more responsive and much prettier).
- Originally developed ~2009 to replace the Perl CGI-based Generic Genome Browser (GBrowse) using what is now a pretty outdated JavaScript framework.
- Recently redeveloped in ReactJS and given MUCH more functionality
- Reads many common genomic formats: GFF, VCF, BigWig, BigBed, HiC, CRAM, BAM, PAF, FASTA
- Available as a web app or a desktop application



Old JBrowse



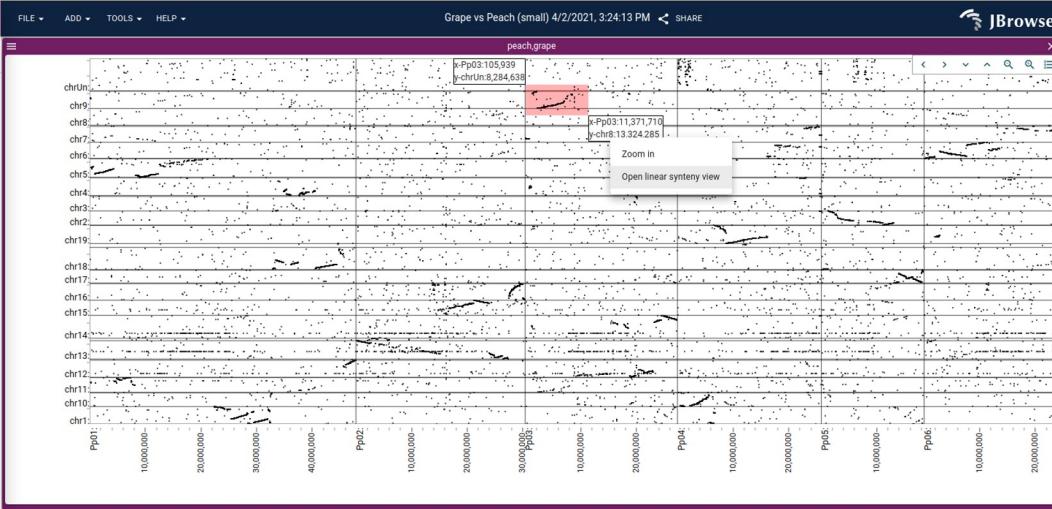
New functionality/views in JBrowse 2

- Visualize comparative genome data with dot plot and synteny views
- Visualize structural variation with circular views and “split” linear genome views
- Use a spreadsheet view to sort and filter features and then click to get context in linear genome views.
- JEXL (JavaScript EXpression Language) allows both admins and users to create callbacks that modify views.
- Plugins written in ReactJS allow a lot of customization: new view types, new track types, new JEXL functions

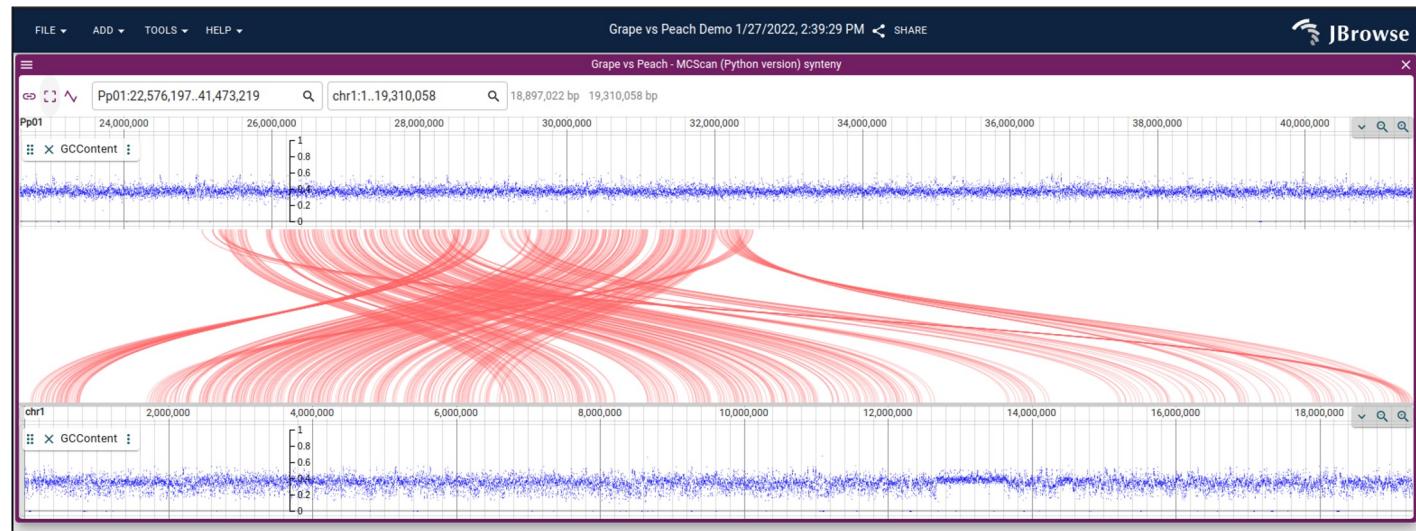


Comparative Genomics

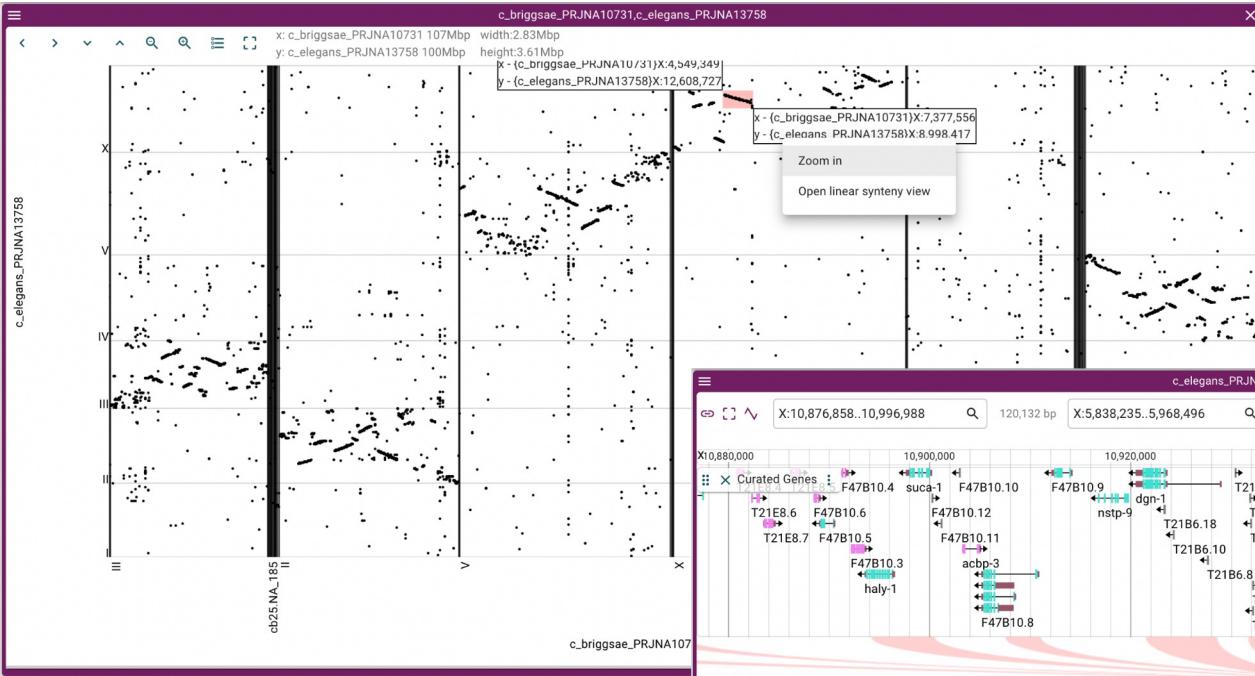
Synteny view



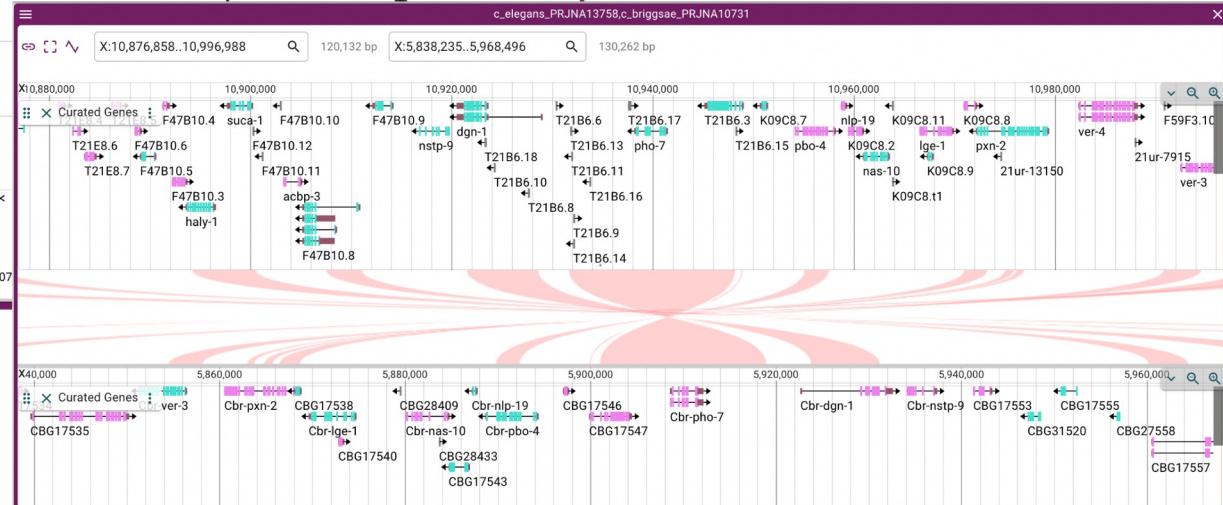
Dotplot view



Dotplot and synteny views



Dotplot comparing *C. elegans* and *C. briggsae* genomes

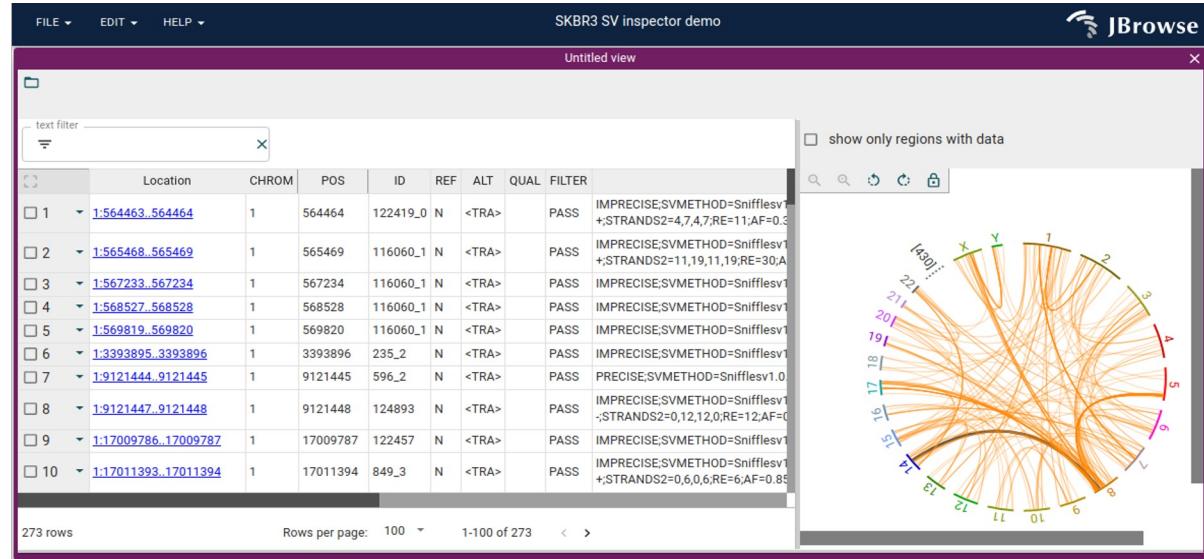


Synteny view showing about 100kb of *C. elegans* and *C. briggsae* genomes, showing an inversion.



The Structural Variant Inspector

- Provide JBrowse with a tabular file of structural variants and it will produce an interactive series of views for analysis
- Clicking on a chord on the circular view opens a breakpoint split view

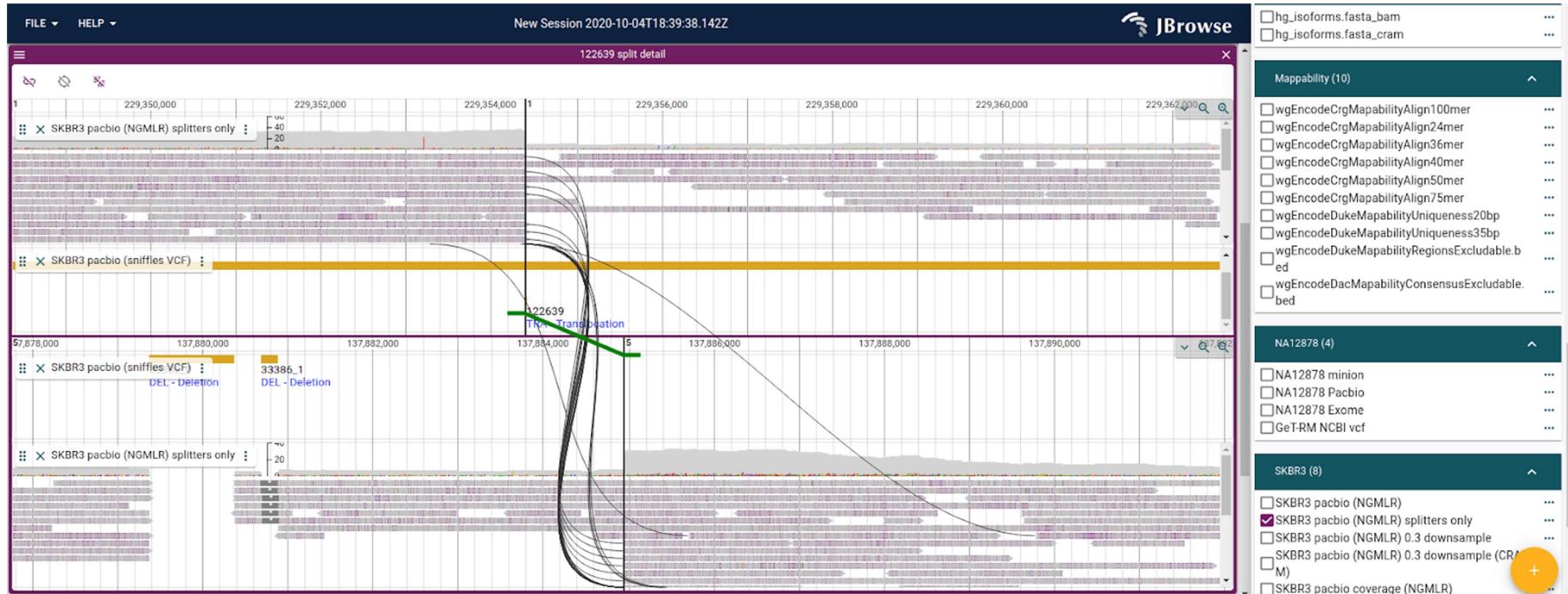


A core component of the SV inspector is the joint spreadsheet and circular views that assist the user in navigating their list of SV's





The Structural Variant Inspector



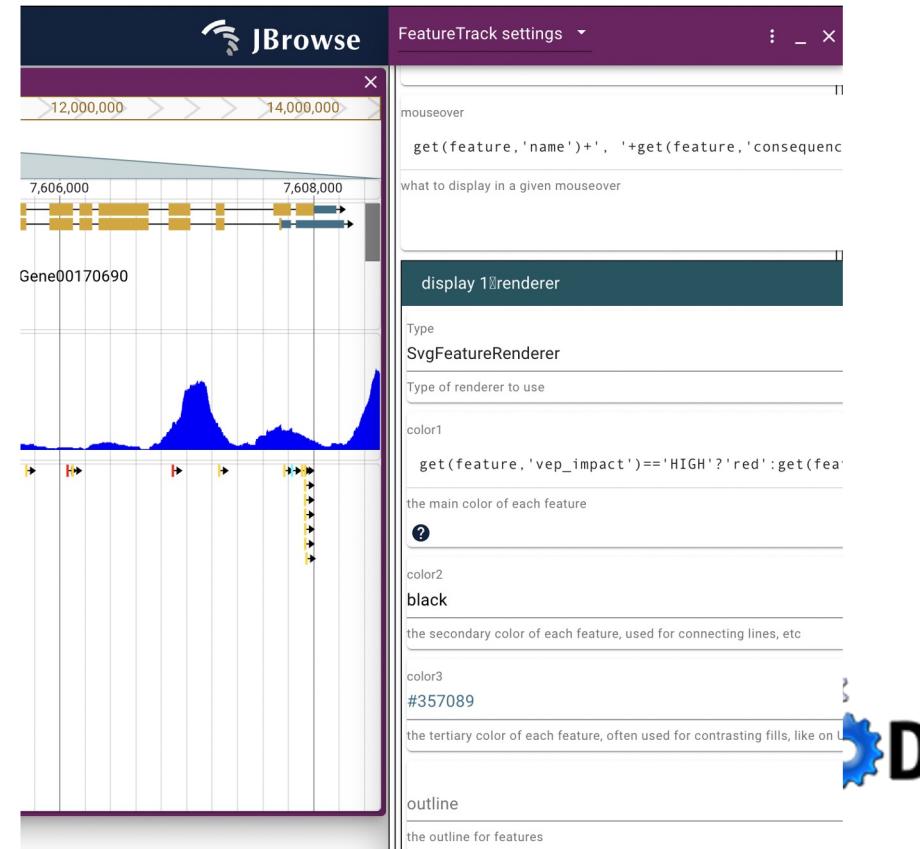
The Breakpoint split view will clearly illustrate your variant (shown: a translocation in the SKBR3 breast cancer cell line)





Graphical configuration editor

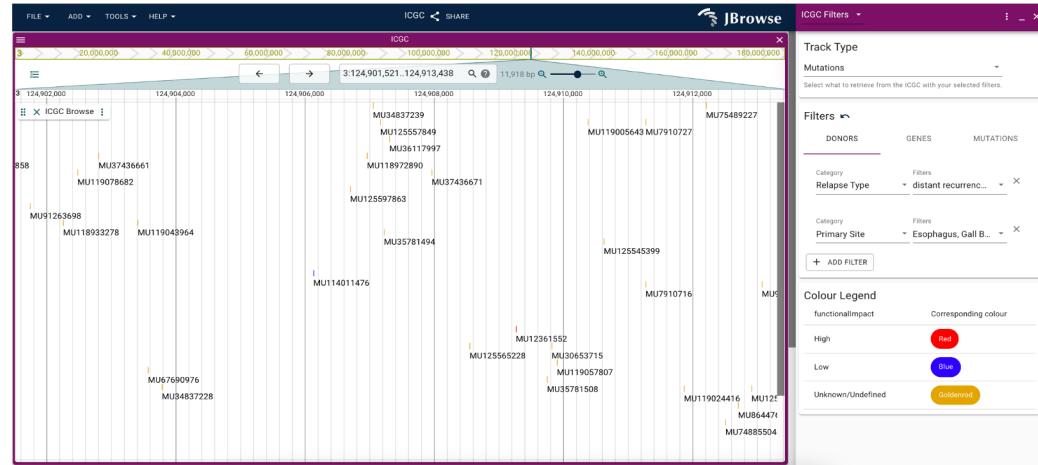
- Both desktop and web apps include a dynamic graphical configuration editor (ie, changes immediately appear in tracks)
 - “Regular” users are restricted to editing personal copies of tracks
- Also, there is a admin editor (called *jbrowse-admin*) that runs a small web server and allows direct editing of the web app config.
- Config uses JEXL instead of JS for security.





Plugin infrastructure

- Allows supplementary feature sets to be added to an instance of JBrowse
- Nearly all core components of JBrowse can be “plugged in” or extended upon through plugins
- Simple development steps to create your own third party plugin for JBrowse 2



Screenshot of the ICGC plugin: a plugin that retrieves and visualizes consortium data within JBrowse 2



JBrowse 2 as an embedded component

- Python, R and JavaScript interfaces exist so that JBrowse can be used with them
- For example, in Jupyter Notebooks
- Or in webpages that use them



File Edit View Run Kernel Tabs Settings Help

+ ☰ ↻ 🔍 Filter files by name

Name Last Modified

- Carsonella... 4 days ago
- Finding rar... 4 days ago
- Haemophil... 4 days ago
- LICENSE 4 days ago
- README.md 4 days ago

Haemophilus influenzae_JE + Python 3 (ipykernel)

```
config.add_df_track(skew_df,'Cumulative GC skew',track_id='df_gc_skew',overwrite=True)
```

Finally, we set a few things about how we want the initial view of JBrowse to look (location, tracks that are turned on) and then launch the genome browser. Note that as we zoom, it might get a little "jerky," as the config file is fairly large, since all of the data we created above is in the config. If we wanted to make this into something we showed other people, we'd want to create data files for each of these tracks.

```
[20]: location = refseq_name + ":1000..3500"
print(location)
```

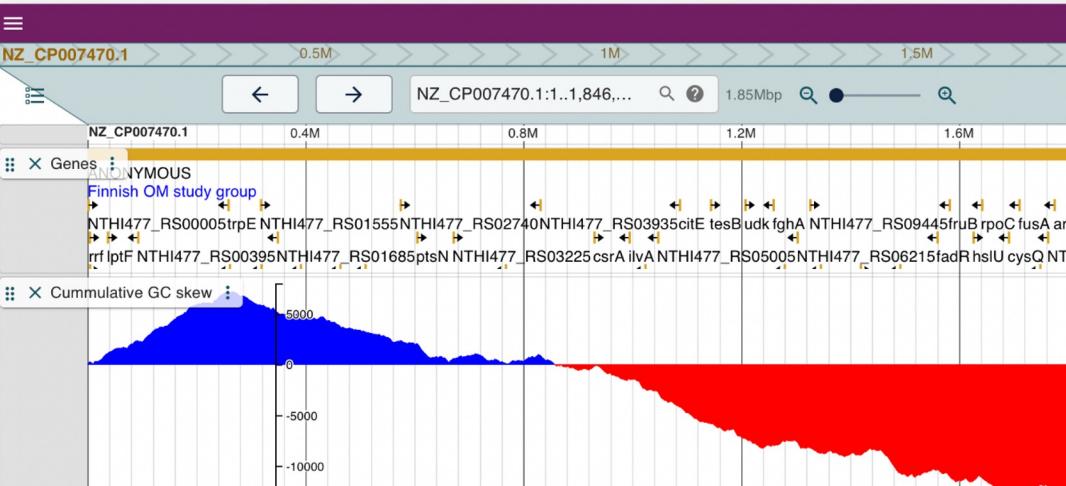
NZ_CP007470.1:1000..3500

```
[21]: config.set_location(location)
#config.set_default_session(['Genes','df_gc_skew','df_percent_gc','df_gc_asym'], False)
config.set_default_session(['Genes','df_gc_skew'], False)
full_conf = config.get_config()
launch(full_conf, port=3003, height=600) #height sets the height of the iframe containing JBrowse, but
# JBrowse sets the height of itself. To use all of the height allocated,
# you have to manually drag the bottom edge of jbrowse down.
```

NZ_CP007470.1 0.5M 1M 1.5M

Genes ANONYMOUS Finnish OM study group NTHI477_RS00005trpE NTHI477_RS01555NTHI477_RS02740NTHI477_RS03935citE tesBudk fghA NTHI477_RS09445fruB rpoC fusA aroE rrif lptF NTHI477_RS00395NTHI477_RS01685ptsN NTHI477_RS03225csrA ilvANTHI477_RS05005NTHI477_RS06215fadRhsIU cysQ NTHI4

Cumulative GC skew





▼ Page Content

Overview

Expression

External Links

Gene Ontology

Genetics

History

Homology

Human Diseases

Interactions

Location

Mapping data

Pathways

Phenotypes

Reagent

References

Sequence features

Sequences

▼ Tools

Tree Display

Genetic Map

Nucl. Aligner

▼ My WormBase

My Favorites

My Library

▼ Recent Activity

turn on history

history logging is off

Comments (0)

▼ Overview

Served from Datomic

etr-1 (ELAV-Type RNA binding-protein family)

Predicted to enable RNA binding activity. Involved in determination of adult lifespan. Located in nucleus. Expressed in body wall musculature; cloacal sphincter muscle; copulatory spicule; non-striped muscle; and somatic gonad. Human ortholog(s) of this gene implicated in developmental and epileptic encephalopathy. Is an ortholog of human CELF1 (CUGBP Elav-like family member 1).

evidence

» Legacy manual gene description

» Also refers to

» Curatorial remarks

Species: *Caenorhabditis elegans*
Sequence: T01D1.2

Other name: CELE_T01D1.2

Type: protein coding

Gene class: etr

Status: Live

Clone: T01D1

Parent seq: CHROMOSOME_II

Gene name Jonathan Hodgkin
evidence:

Comparative Integrated model

Info: organism details available at the



WormBase ID: WBGene00001340

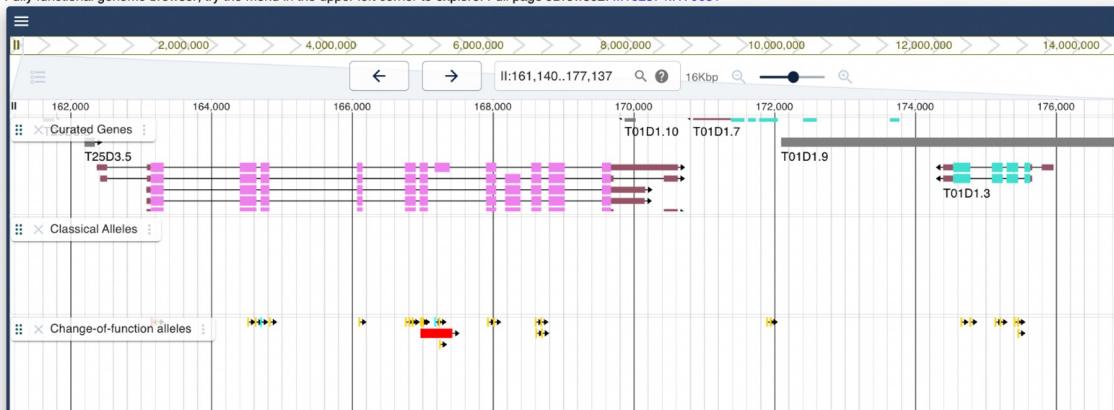
▼ Location

Served from Datomic

Genetic position: II:-15.90 +/- 0.014 cM

Genomic position: II:162374..170631 (Legacy Genome Browser: II:162374..170631)

JBrowse 2 Beta: Fully functional genome browser; try the menu in the upper left corner to explore. Full page JBrowse2: II:162374..170631



Embedded JBrowse is essentially fully featured

- Embedded JBrowse works as a complete genome browser
- You can add/remove tracks
- Pan/zoom
- Goto other locations



JBrowse 2 resources

- <http://jbrowse.org/>
- <https://github.com/GMOD/jbrowse-components> - primary repo
- <https://github.com/GMOD> - many other JS repos
- <https://github.com/GMOD/jbrowse-jupyter>
- <https://github.com/scottcain/jbrowse-jupyter-examples>
- <https://gitter.im/GMOD/jbrowse>



Intro to Apollo

- Apollo is a graphical genome feature structure editor (ie, so you can create and modify features like genes, transcripts, etc)
- First incarnation was a Java desktop application written by FlyBase in the early 2000s (“Apollo 1”).
- Rewritten as a JBrowse 1 plugin several years ago (“WebApollo” and “Apollo 2”)
- Currently being rewritten as a JBrowse 2 plugin (“Apollo 3”)
- Apollo 2 and 3 are both live and interactive like Google Docs (so “gene editing wars” are possible)

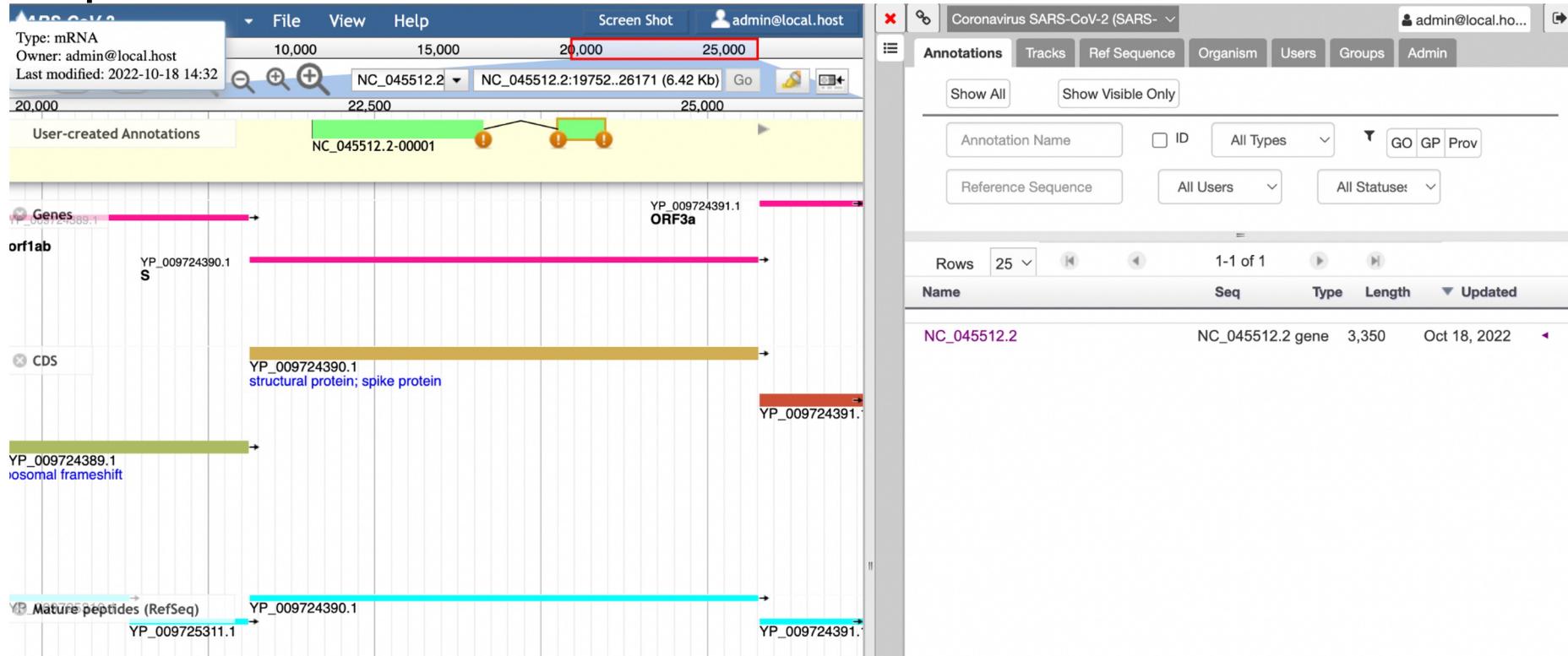


Who uses Apollo?

- Professional curators at model organism databases
- “Casual” users who are interested in “their” gene or gene family
- “Annotation Jamboree” users (locked in a room with a group for a few days to annotate a genome)
- Students learning about genomics



Apollo user interface



Apollo resources

- Docs: <https://genomearchitect.readthedocs.io/en/latest/>
- Source: <https://github.com/GMOD/Apollo/>

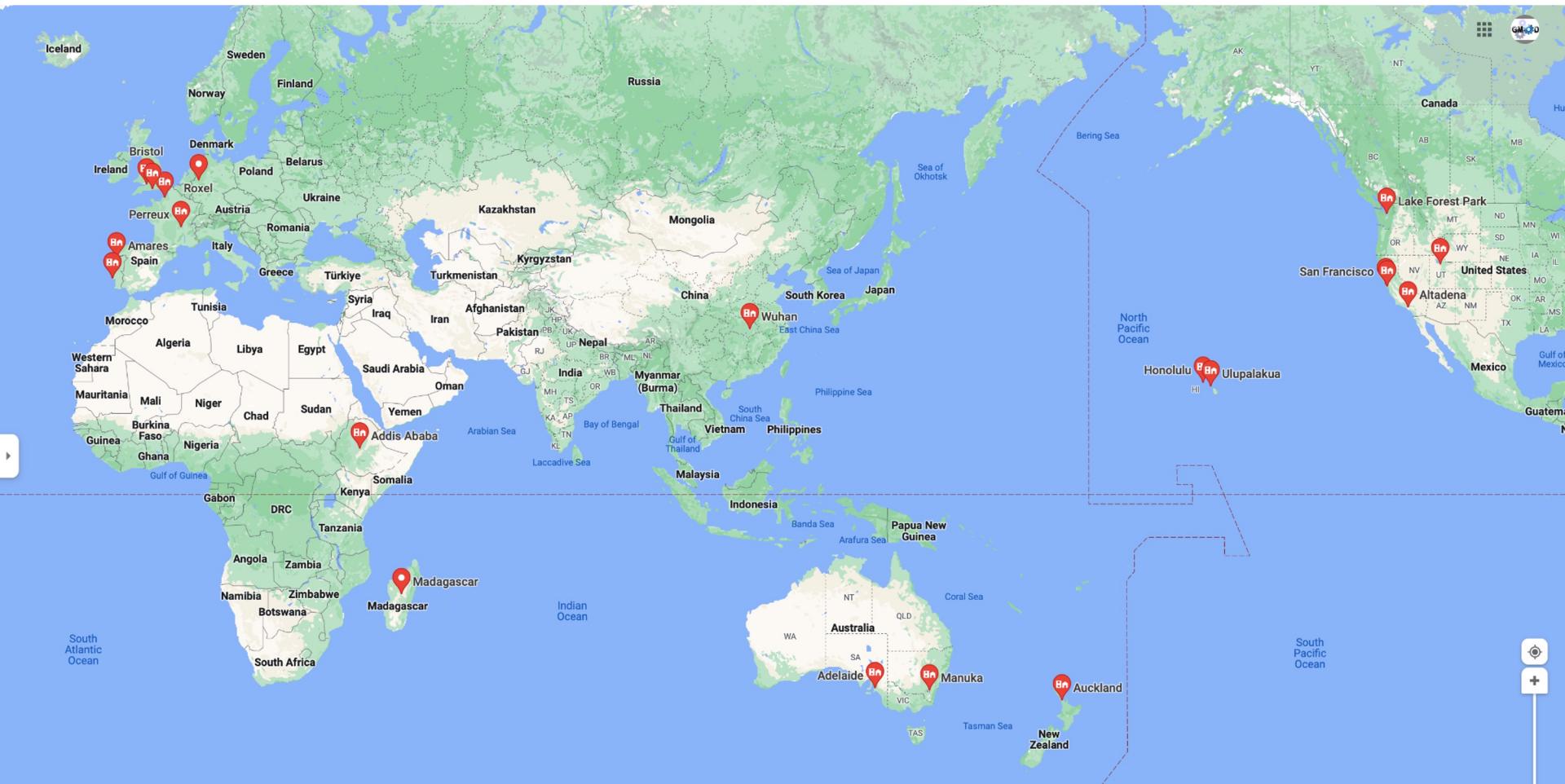


A little about a “side” project I’m working on

Building a web portal for ...



World Wide Worms (WWW)!



Err, what?

- Cristian Riccio, researcher from Cambridge University, contacted WormBase with a proposal:
 - He was wrapping up his PhD dissertation that involved natural diversity of *C. elegans*
 - He had 20 fully assembled genomes from worms around the world
 - And felt that other worm researchers would be interested in this data set
- The feeling I get is that most of the people at WormBase had the response along the lines of “Maybe...?”
- Then it landed in my lap, and a plan began to form.



Build a stand alone portal

- There isn't really a good way to incorporate data like this into "WormBase proper"
- That's part of why my colleagues at WB weren't exactly sure what to do with this data
- As an alternative, I suggested a portal that was either completely or mostly a JBrowse instance with comparative data between the various strains



Never underestimate the bandwidth...

of a station wagon full of tapes hurtling down the highway. ~~Andrew S. Tannenbaum

So I got a hard drive with (just) a few hundred GB of annotations, assemblies and mapping data sent via post from Switzerland.



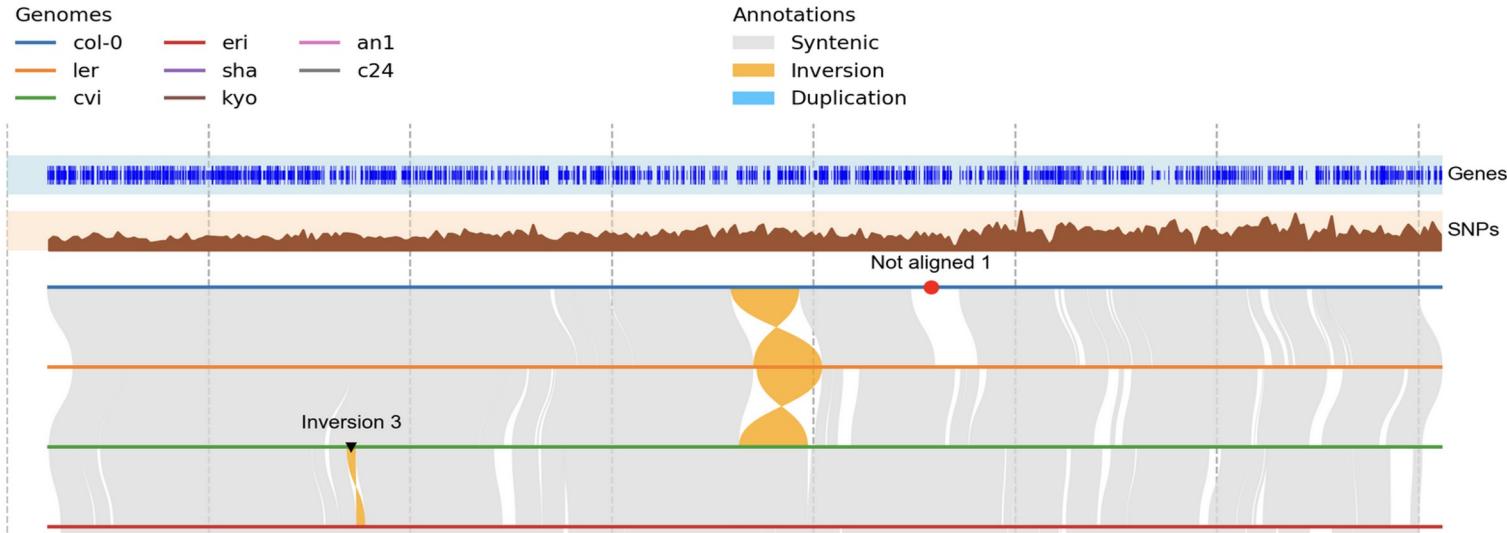
How to best guide users to things that might be interesting?

- Of course, users could search for a gene and compare that between different strains, but most of the time, those are going to be exactly or nearly exactly the same between them
- Find the globally visible differences between the assemblies.
- Already familiar with minimap2 (Heng Li) (JBrowse natively displays results from it) but it didn't quite do the trick visually for largely identical assemblies.



Found SyRI and PlotSR

- Compares alignments between two chromosome-level assemblies and identifies synteny and structural rearrangements.
- <https://github.com/schneebergerlab/syri>
- Goel, M., Sun, H., Jiao, W. et al. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* 20, 277 (2019) doi:10.1186/s13059-019-1911-0



Imagine multiple slides here

Of me complaining about the difficulties of setting up python applications to run inside of Docker containers. Suffice it to say, I was quite whiny but got it working in the end.



What the entrypoint script does

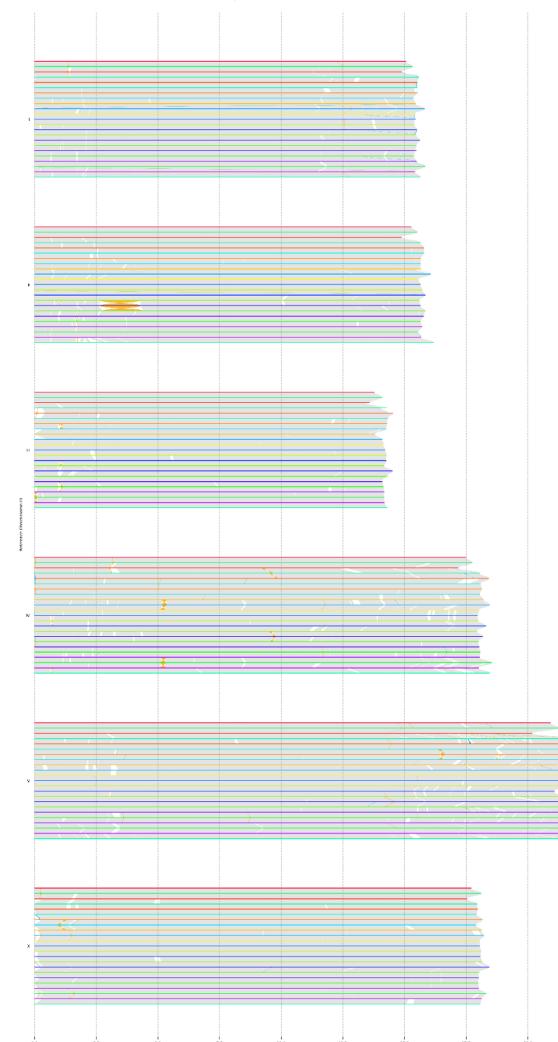
- Run “serial pairwise” minimap2 (A vs B, B vs C, C vs D, etc) (n-1 executions)
- Activate conda
- Run syri on the minimap2 output for each pair (n-1 executions)
- Run plotsr once on all of the syri outputs.



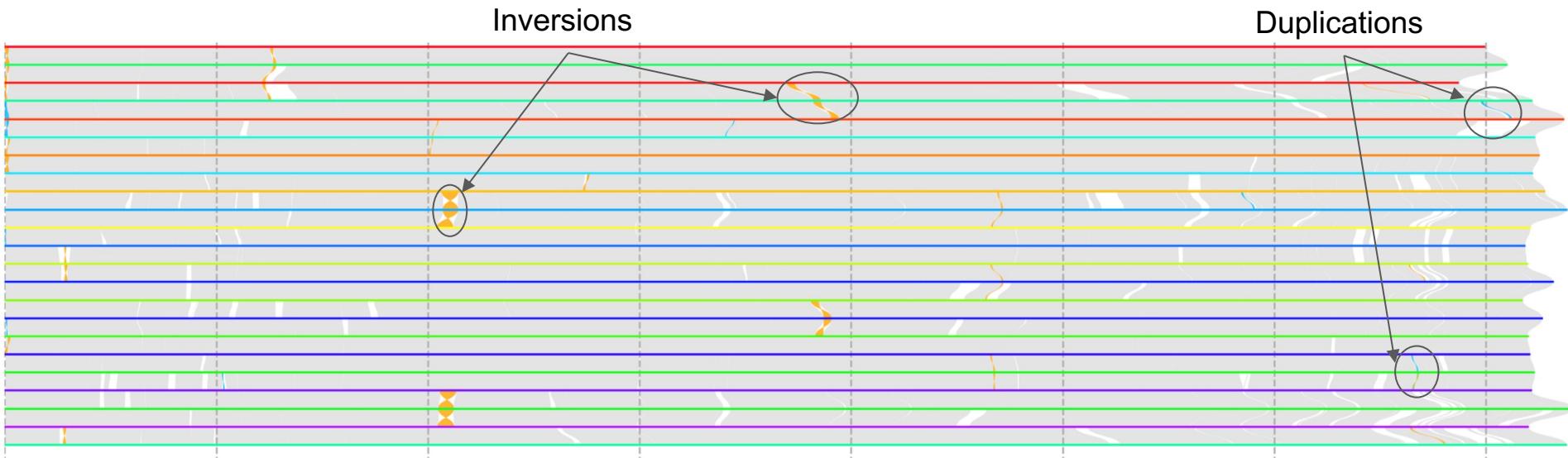
Tall output

The result is all 21 assemblies stacked by chromosome (I through V, plus X)

Output can be SVG, PDF or PNG



plotsr output (chromosome II)



Probably also translocations and deletions in here (hard to tell what's going on in the gaps)



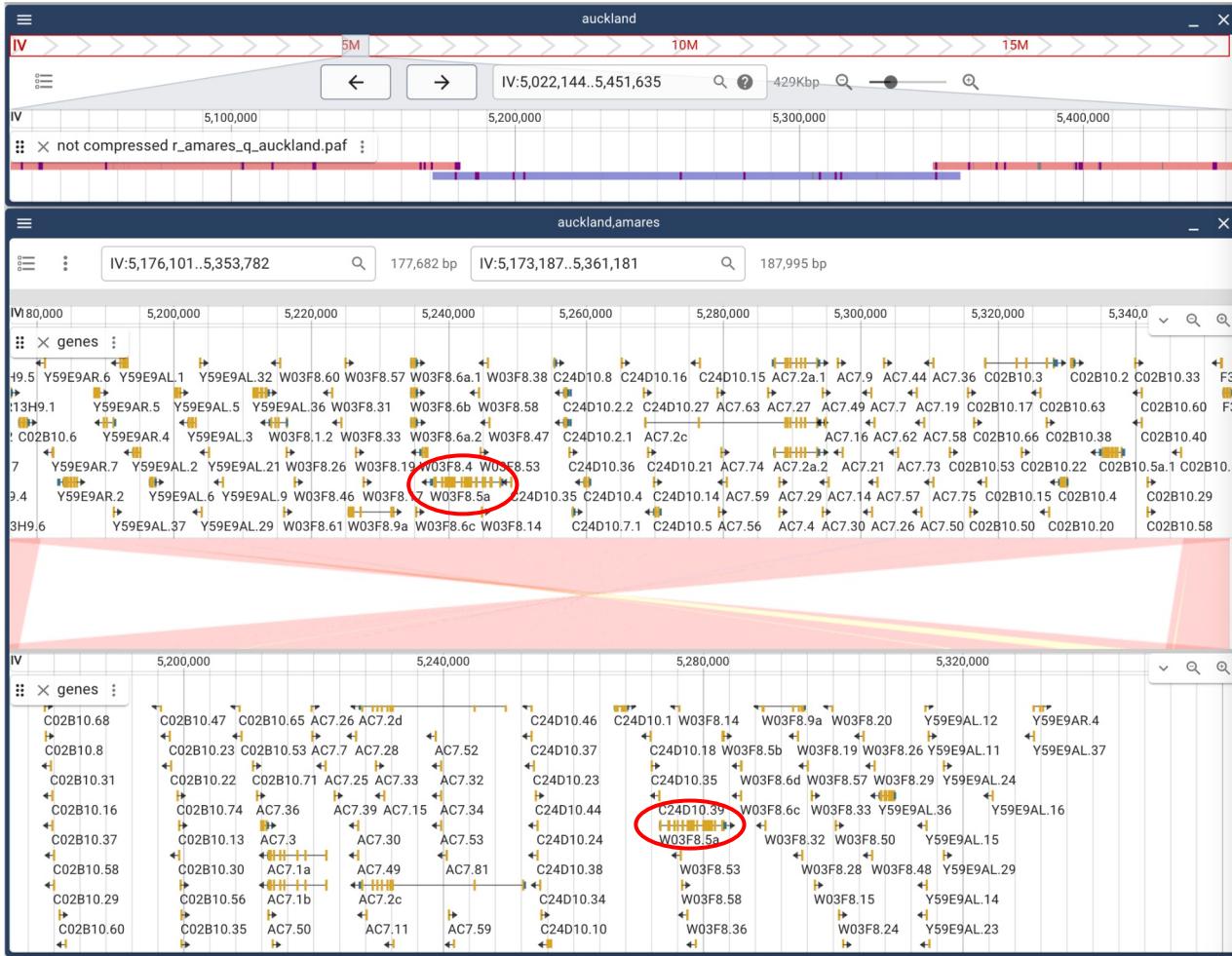
Idea: Break into per-chromosome images and make them into image maps



(That is, a set of defined coordinates on an image that, when clicked on, go to a specific (JBrowse) link)



To open a view like this



So what does the “portal” look like (in my head)?

- An index.html page that has a little information and six image maps
- A JBrowse instance that has 21 assemblies, each with:
 - a set of gene annotations
 - four minimap2 outputs (to the ones before and after in the series, and to the main references)
 - (Possibly—still trying to decide if useful) BAM alignments from the approx 250x coverage
- Also in JBrowse: the main reference sequences (N2 and VC2010) with VCFs for each of the strain assemblies and (why not) all of the data that they have at WormBase.



Wrap up/Demos

I can demo most things—I'll start with JBrowse 2, but please speak up if there something else you'd like to see.

Lots o'worms: <https://riccio.d2jjb0xowet5mr.amplifyapp.com/>

Structural variation:

https://jbrowse.org/code/jb2/v2.7.1/?config=test_data%2Fconfig_demo.json&session=local-DwmlfSK0XZfpixc87RSMP

Alliance variants: <https://stage.dgaayxgqoarxf.amplifyapp.com/?session=local-E6xoJTWzfodeh3IHNse6d>



Acknowledgements

People who contributed slides to make my life easier:

- Lacey Ann Sanderson (Tripal)
- Carson Holt (MAKER)
- Beatriz Serrano-Solano (Galaxy)

The literally hundreds (thousands?) of people in the GMOD community who have contributed in so many roles (developers, curators, researchers, documenters, etc).

Finally, Lincoln Stein (co-originator of this course) for supporting me and the GMOD project even when direct funding ended.



Resources

<http://gmod.org/>

<http://jbrowse.org/>

<http://tripal.info/>

<http://usegalaxy.org/>

