JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

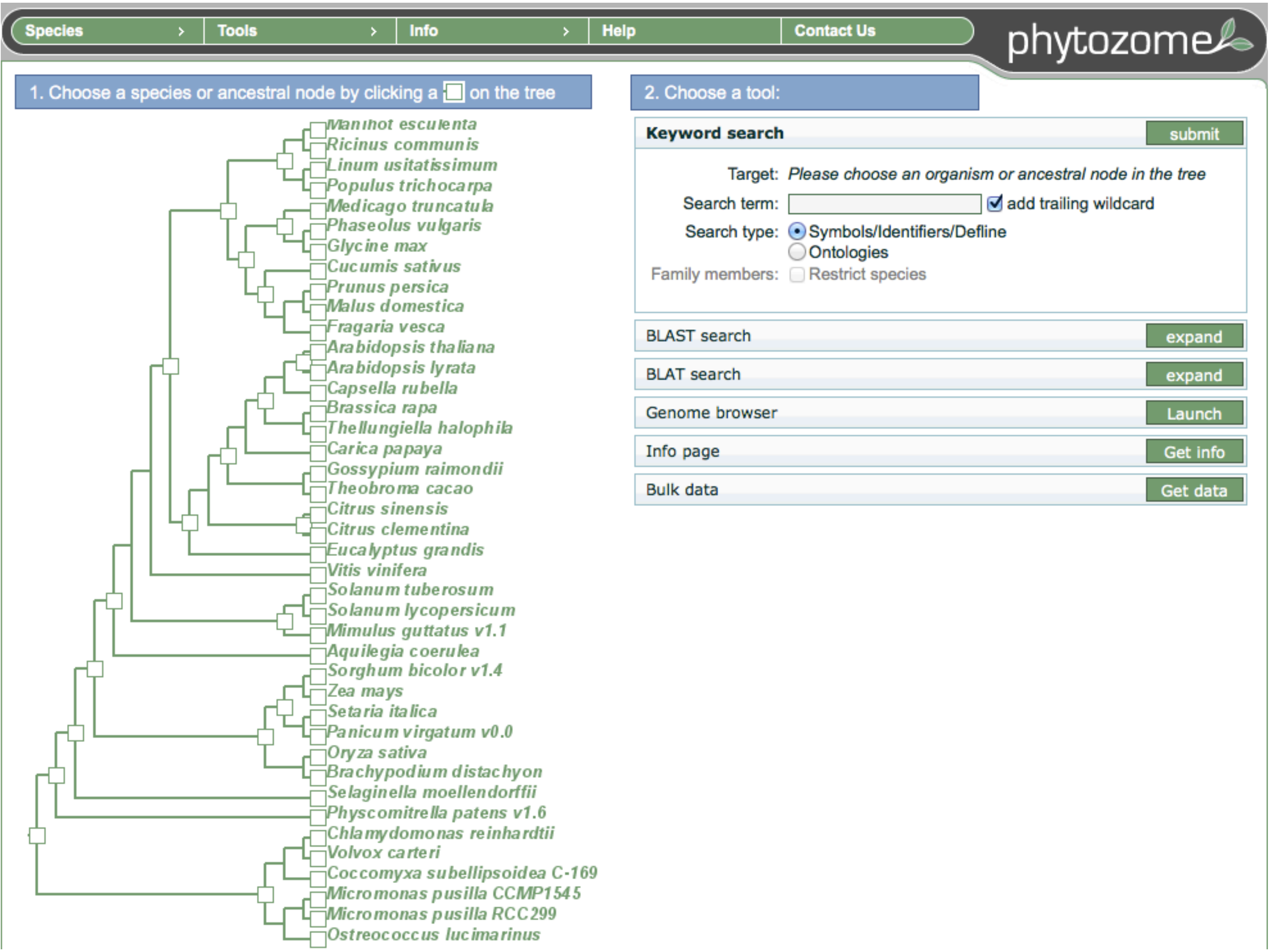
Migrating from GBrowse to JBrowse

Richard D. Hayes
Plant Genomics Group
www.phytozome.net

Overview

- Phytozome Introduction
- Track Data Conversion
- Name Indexing
- Diversity Data
- Expression Data
- Summary

Current Phytozome v9



1. Choose a species or ancestral node by clicking a □ on the tree

2. Choose a tool:

Keyword search submit

Target: *Please choose an organism or ancestral node in the tree*

Search term: ☑ add trailing wildcard

Search type: ● Symbols/Identifiers/Defline
○ Ontologies

Family members: □ Restrict species

BLAST search expand

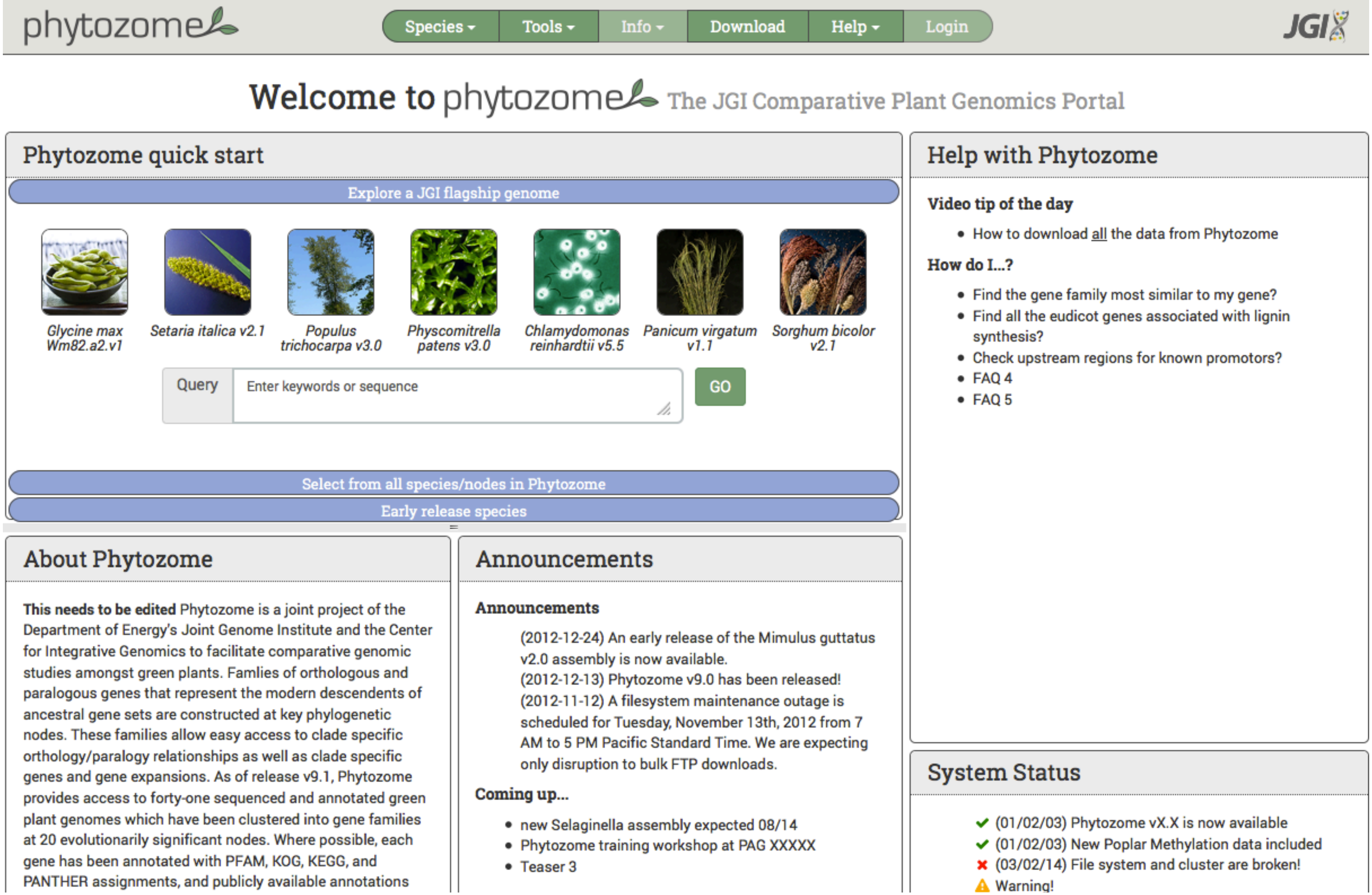
BLAT search expand

Genome browser Launch

Info page Get info

Bulk data Get data

Phytozome: Plant Comparative Genomics Portal



Welcome to phytozome The JGI Comparative Plant Genomics Portal

Phytozome quick start

Help with Phytozome

Video tip of the day

- How to download all the data from Phytozome

How do I...?

- Find the gene family most similar to my gene?
- Find all the eudicot genes associated with lignin synthesis?
- Check upstream regions for known promotors?
- FAQ 4
- FAQ 5

About Phytozome

This needs to be edited Phytozome is a joint project of the Department of Energy's Joint Genome Institute and the Center for Integrative Genomics to facilitate comparative genomic studies amongst green plants. Famlies of orthologous and paralogous genes that represent the modern descendents of ancestral gene sets are constructed at key phylogenetic nodes. These families allow easy access to clade specific orthology/paralogy relationships as well as clade specific genes and gene expansions. As of release v9.1, Phytozome provides access to forty-one sequenced and annotated green plant genomes which have been clustered into gene families at 20 evolutionarily significant nodes. Where possible, each gene has been annotated with PFAM, KOG, KEGG, and PANTHER assignments, and publicly available annotations

Announcements

Announcements

(2012-12-24) An early release of the Mimulus guttatus v2.0 assembly is now available.
(2012-12-13) Phytozome v9.0 has been released!
(2012-11-12) A filesystem maintenance outage is scheduled for Tuesday, November 13th, 2012 from 7 AM to 5 PM Pacific Standard Time. We are expecting only disruption to bulk FTP downloads.

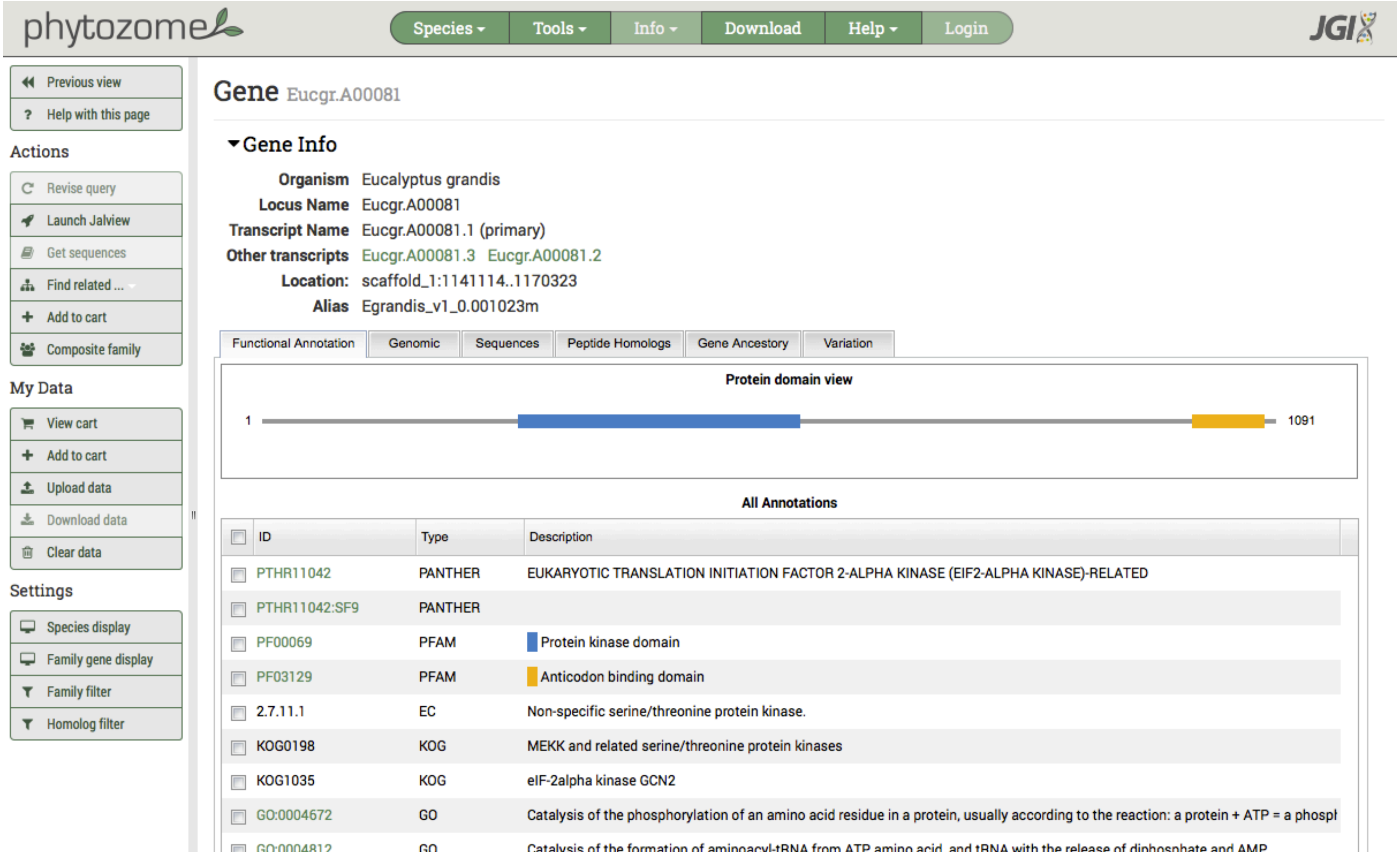
Coming up...

- new Selaginella assembly expected 08/14
- Phytozome training workshop at PAG XXXXX
- Teaser 3

System Status

✓ (01/02/03) Phytozome vX.X is now available
✓ (01/02/03) New Poplar Methylation data included
✗ (03/02/14) File system and cluster are broken!
⚠ Warning!

Phytozome: Plant Comparative Genomics Portal



Gene Eucgr.A00081

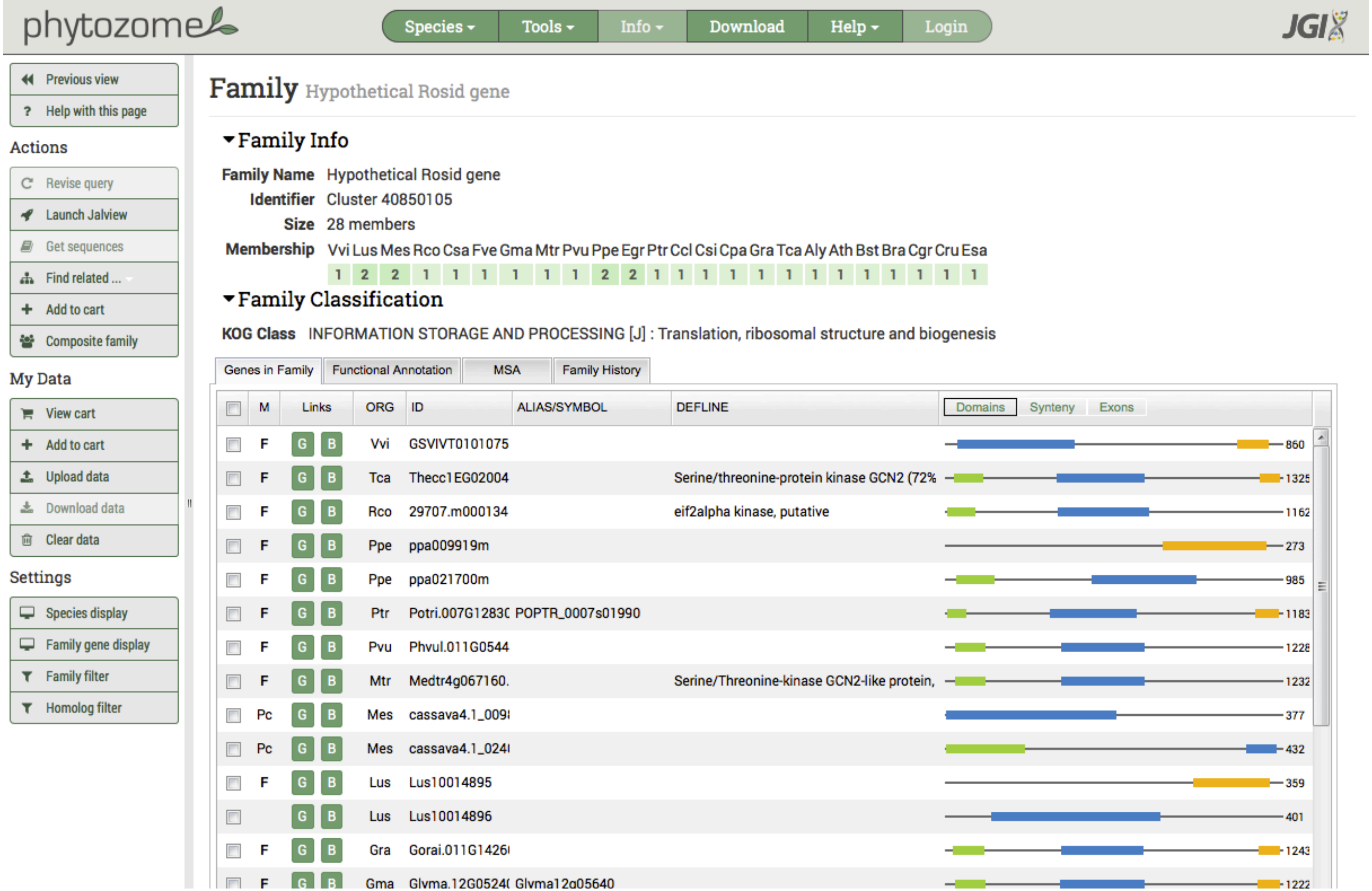
Gene Info

Organism	Eucalyptus grandis
Locus Name	Eucgr.A00081
Transcript Name	Eucgr.A00081.1 (primary)
Other transcripts	Eucgr.A00081.3 Eucgr.A00081.2
Location:	scaffold_1:1141114..1170323
Alias	Egrandis_v1_0.001023m

All Annotations

□	ID	Type	Description
□	PTHR11042	PANTHER	EUKARYOTIC TRANSLATION INITIATION FACTOR 2-ALPHA KINASE (EIF2-ALPHA KINASE)-RELATED
□	PTHR11042:SF9	PANTHER	
□	PF00069	PFAM	🟦 Protein kinase domain
□	PF03129	PFAM	🟧 Anticodon binding domain
□	2.7.11.1	EC	Non-specific serine/threonine protein kinase.
□	KOG0198	KOG	MEKK and related serine/threonine protein kinases
□	KOG1035	KOG	eIF-2alpha kinase GCN2
□	GO:0004672	GO	Catalysis of the phosphorylation of an amino acid residue in a protein, usually according to the reaction: a protein + ATP = a phospl
□	GO:0004812	GO	Catalysis of the formation of aminoacyl-tRNA from ATP amino acid, and tRNA with the release of diphosphate and AMP

Phytozome: Plant Comparative Genomics Portal



Family Hypothetical Rosid gene

▾Family Info

Family Name Hypothetical Rosid gene

Identifier Cluster 40850105

Size 28 members

Membership

Vvi	Lus	Mes	Rco	Csa	Fve	Gma	Mtr	Pvu	Ppe	Egr	Ptr	Ccl	Csi	Cpa	Gra	Tca	Aly	Ath	Bst	Bra	Cgr	Cru	Esa
1	2	2	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1

▾Family Classification

KOG Class INFORMATION STORAGE AND PROCESSING [J] : Translation, ribosomal structure and biogenesis

□	M	Links	ORG	ID	ALIAS/SYMBOL	DEFLINE	Domains Synteny Exons
□	F	G B	Vvi	GSVIVT0101075			860
□	F	G B	Tca	Thecc1EG02004		Serine/threonine-protein kinase GCN2 (72%	1325
□	F	G B	Rco	29707.m000134		eif2alpha kinase, putative	1162
□	F	G B	Ppe	ppa009919m			273
□	F	G B	Ppe	ppa021700m			985
□	F	G B	Ptr	Potri.007G1283(	POPTR_0007s01990		1183
□	F	G B	Pvu	Phvul.011G0544			1228
□	F	G B	Mtr	Medtr4g067160.		Serine/Threonine-kinase GCN2-like protein,	1232
□	Pc	G B	Mes	cassava4.1_009			377
□	Pc	G B	Mes	cassava4.1_024			432
□	F	G B	Lus	Lus10014895			359
□		G B	Lus	Lus10014896			401
□	F	G B	Gra	Gorai.011G1426			1243
□	F	G B	Gma	Glyma.12G0524(	Glyma12g05640		1222

- Phytozome Introduction
- **Track Data Conversion**
- Name Indexing
- Diversity Data
- Expression Data
- Summary

Need For Parallel Data Processing

45 genome annotations, average 8 data tracks

- 5 or 6 for some Chlorophytes
- *Chlamydomonas reinhardtii* has 27

Large variation in track feature density

Populus trichocarpa

Primary Tr: 41,335	Alt. Tr: 31,678
EST Alignments: 3,898,010	EST Assemblies: 237,993
BLASTX Proteins: 2,237,632	BLATX Proteins: 508,948

Glycine max

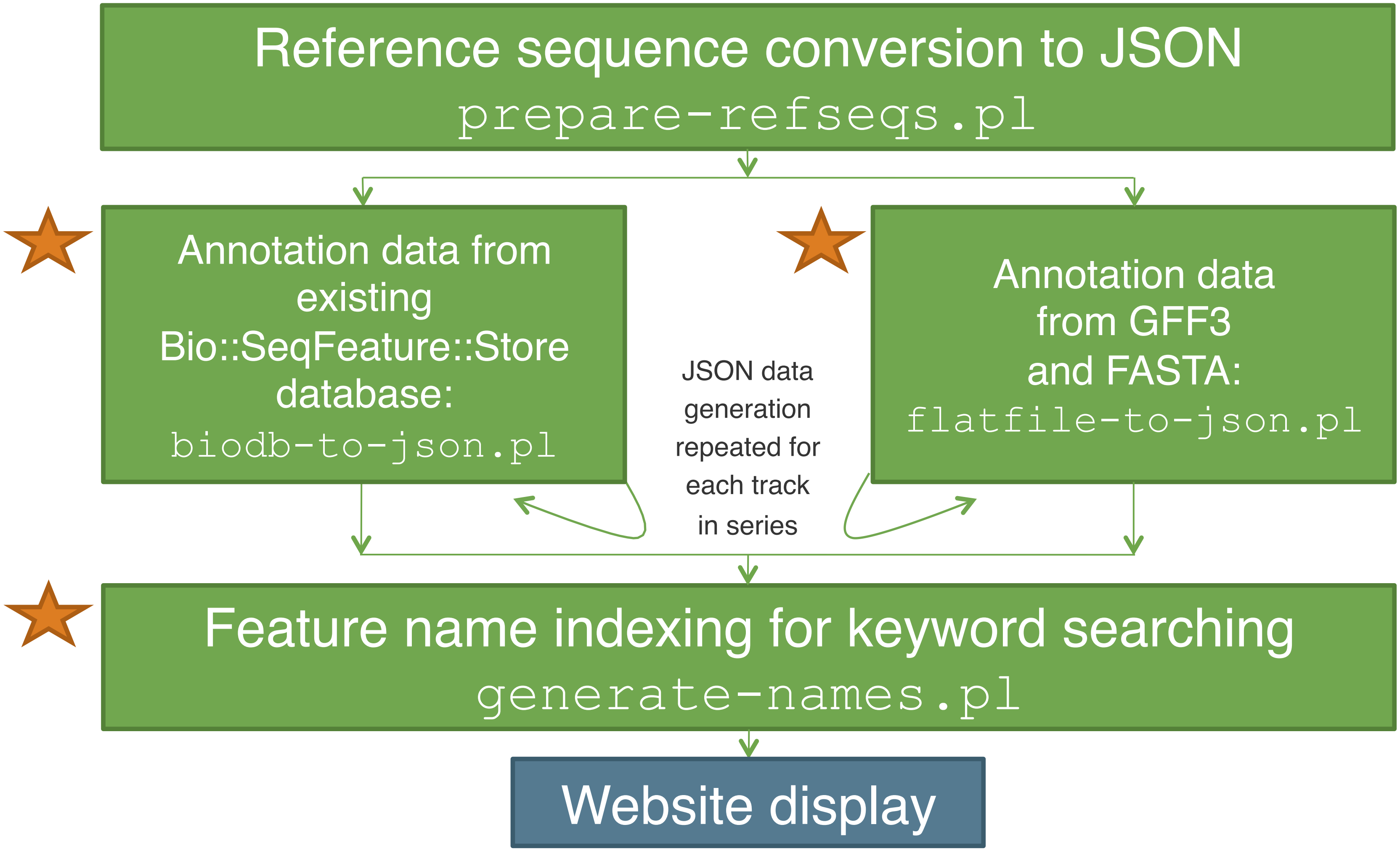
Primary Tr: 56,044	Alt. Tr: 32,603
EST Alignments: 1,974,371	EST Assemblies: 193,245
BLASTX Proteins: 2,710,298	BLATX Proteins: 795,488

NERSC

547 nodes 4680 cores (7.36Gb RAM avg)

“interactive nodes”: 24 core; 256 Gb

Usual Installation Workflow



Reference sequence conversion to JSON
`prepare-refseqs.pl`

Annotation data from existing Bio::SeqFeature::Store database:
`biodb-to-json.pl`

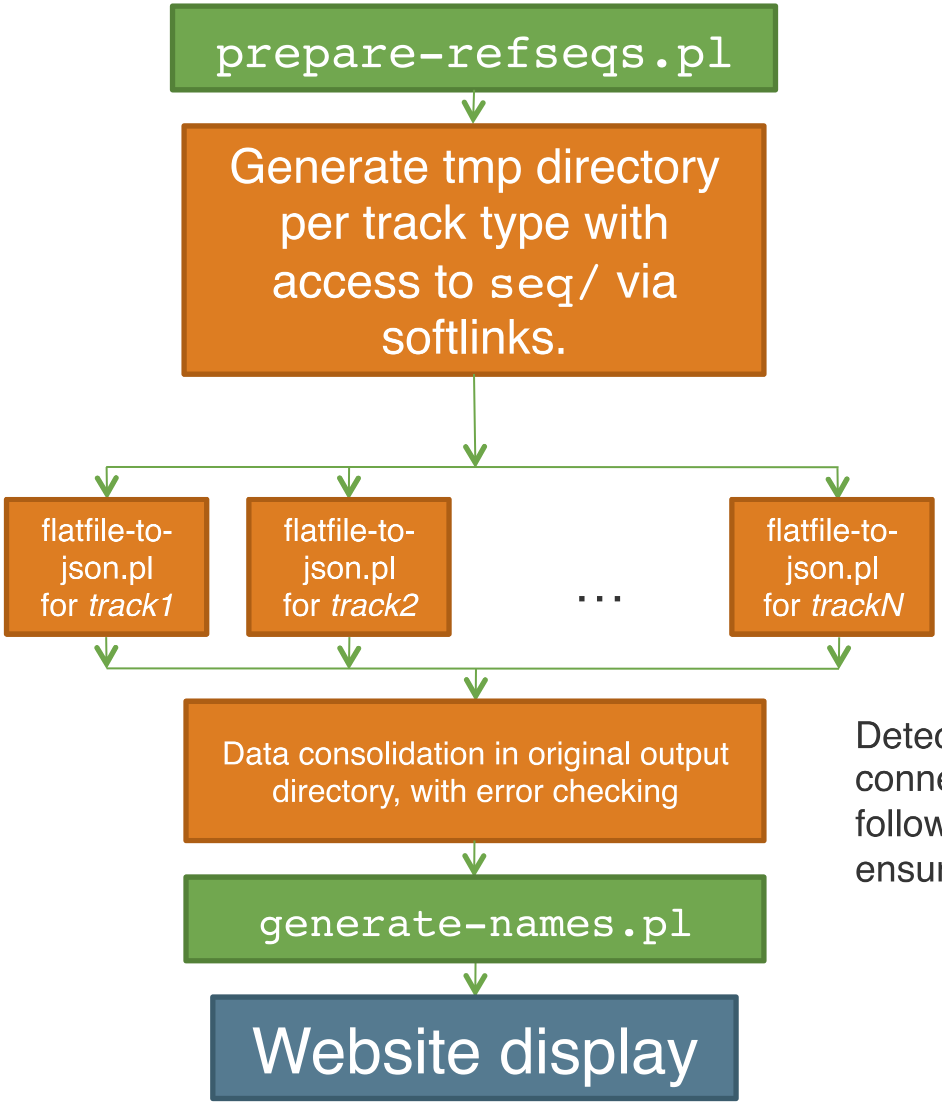
JSON data generation repeated for each track in series

Annotation data from GFF3 and FASTA:
`flatfile-to-json.pl`

Feature name indexing for keyword searching
`generate-names.pl`

Website display

New Parallel Workflow



`prepare-refseqs.pl`

Generate tmp directory per track type with access to `seq/` via softlinks.

flatfile-to-json.pl for *track1*

flatfile-to-json.pl for *track2*

…

flatfile-to-json.pl for *trackN*

Data consolidation in original output directory, with error checking

`generate-names.pl`

Website display

Refseq processing is run as before

e.g. for *P. trichocarpa*:

```
<outroot>/Ptr/
```

↓

```
<outroot>/Ptr.track1.tmp/
<outroot>/Ptr.track2.tmp/
…
<outroot>/Ptr.trackN.tmp/
```

JSON track data generation run in parallel on compute cluster

Detection of potential perl exceptions (e.g. database connection errors, out-of-memory conditions, etc.) followed by JSON data copy and directory diff to ensure a complete data consolidation

biodb Run Time Comparisons

Primary Transcripts	41,335
Alternative Transcripts	31,678
RepeatMasker Masked Regions	573,268
PASA Aligned ESTs	191,633
PASA Assembled ESTs	86,677
PASA Aligned Sibling ESTs	3,706,377
PASA Assembled Sibling ESTs	151,316
BLAT Alignments of ESTs from related species	8,919
Protein alignments by BLASTX (1E-5)	2,237,633
Protein alignments by BLATX (50% ID,20% coverage)	479,524
Assembly Gaps	5,835
v2.2 Annotation Mapped Transcripts	41,763

Populus trichocarpa parent feature counts per track

biodb-to-json.pl runs

Step	User Time	System Time	Wall Clock Time	CPU	Max. vmem RAM Usage
Full data in series	02:39:17	00:08:48	06:15:57	02:48:06	1.600 Gb
Aligned EST SIB (parallel)	01:09:41	00:03:44	03:05:09	03:03:34	5.25 Gb
v2.2 Transcripts (parallel)	00:00:06	00:00:01	00:02:33	00:00:08	127.527 Mb

flatfile Run Time Comparisons

Primary Transcripts	41,335
Alternative Transcripts	31,678
RepeatMasker Masked Regions	573,268
PASA Aligned ESTs	191,633
PASA Assembled ESTs	86,677
PASA Aligned Sibling ESTs	3,706,377
PASA Assembled Sibling ESTs	151,316
BLAT Alignments of ESTs from related species	8,919
Protein alignments by BLASTX (1E-5)	2,237,633
Protein alignments by BLATX (50% ID,20% coverage)	479,524
Assembly Gaps	5,835
v2.2 Annotation Mapped Transcripts	41,763

Populus trichocarpa
parent feature
counts per track

Effectively a 12-fold reduction in run time versus biodb-to-json.pl run in series!

flatfile-to-json.pl runs
(all in parallel)

Step	User Time	System Time	Wall Clock Time	CPU	Max. vmem RAM Usage
v2.2 Transcripts	00:00:04	00:00:00	00:00:19	00:00:05	112.469 Mb
Aligned EST SIB	00:06:42	00:00:04	00:07:28	00:06:46	2.722 Gb
BLASTX Alignments	00:24:59	00:00:16	00:27:38	00:25:16	4.524 Gb

Overview

- Phytozome Introduction
- Track Data Conversion
- **Name Indexing**
- Diversity Data
- Expression Data
- Next Steps

State of Name Indexing

Many algorithmic improvements since JBrowse v1.10.7

We take advantage of several "power user" options in v1.11.0

- Incremental indexing

  Process monitoring and "partial" index checkpointing

  Autocompletion settings for certain tracks where this is useful (Transcripts) and full name indexing where it is not (aligned ESTs)

- Custom search attributes

  Allows for unique links per gene/transcript from Phytozome frontend views via our internal DB ids

  Crucial for genomes where prior versions are mapped forward, such that transcript name are no longer guaranteed to be unique themselves

Overview

- Phytozome Introduction
- Track Data Conversion
- Name Indexing
- **Diversity Data**
- Expression Data
- Next Steps

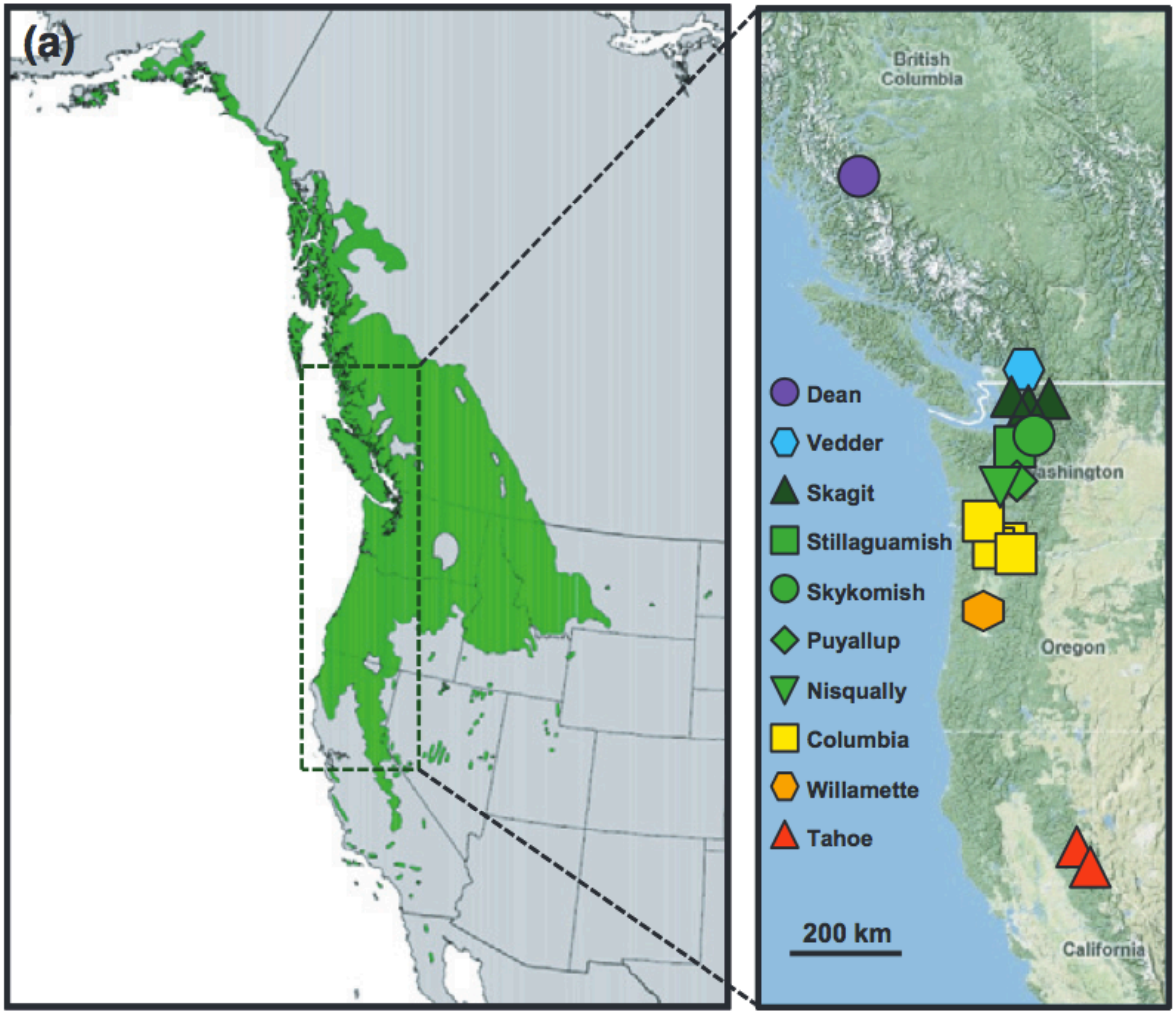
One of Our Fastest Growing Data Types

Initial project started with resequencing 16 *P. trichocarpa* individuals

Joint variant calling (GATK)

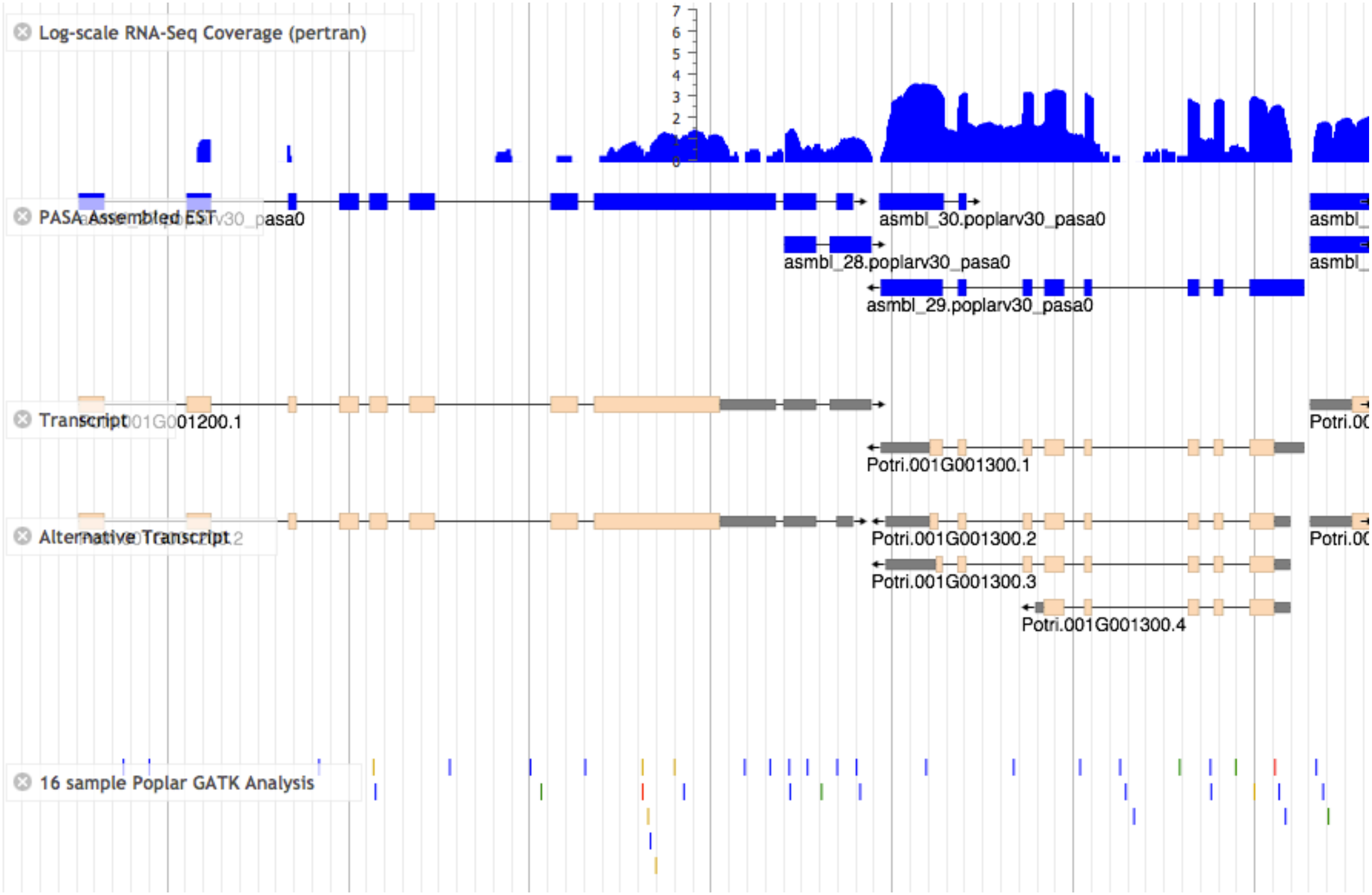
6,717,307 SNV
679,681 deletions
793,463 insertions
74,085 indels

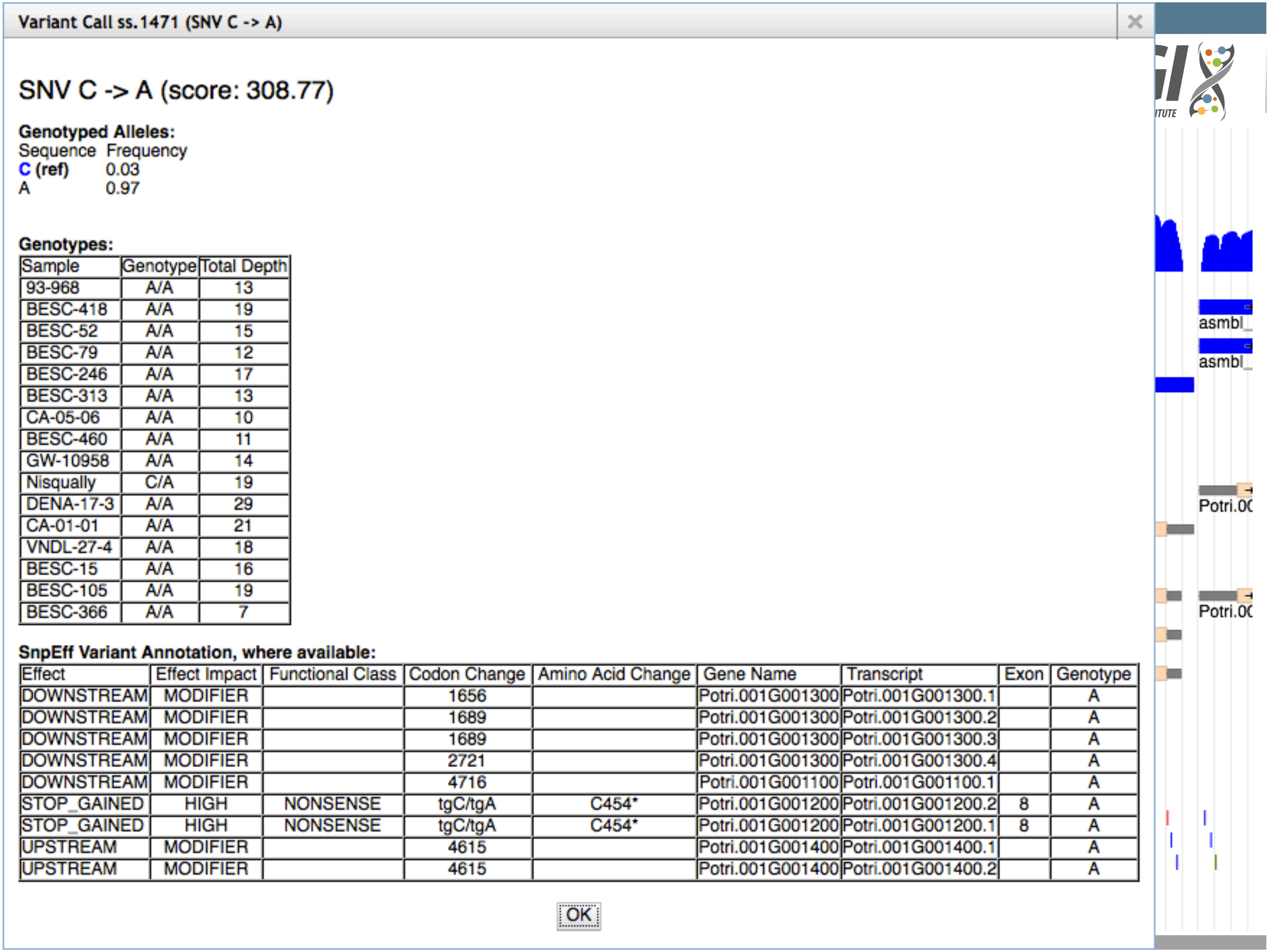
Soon to expand to over 1000 trees



Slavov, GT *et al.* (2012) New Phytologist, 196: 713–725.

Direct VCF Visualization is Nifty





Variant Call ss.1471 (SNV C -> A)

SNV C -> A (score: 308.77)

Genotyped Alleles:

Sequence	Frequency
C (ref)	0.03
A	0.97

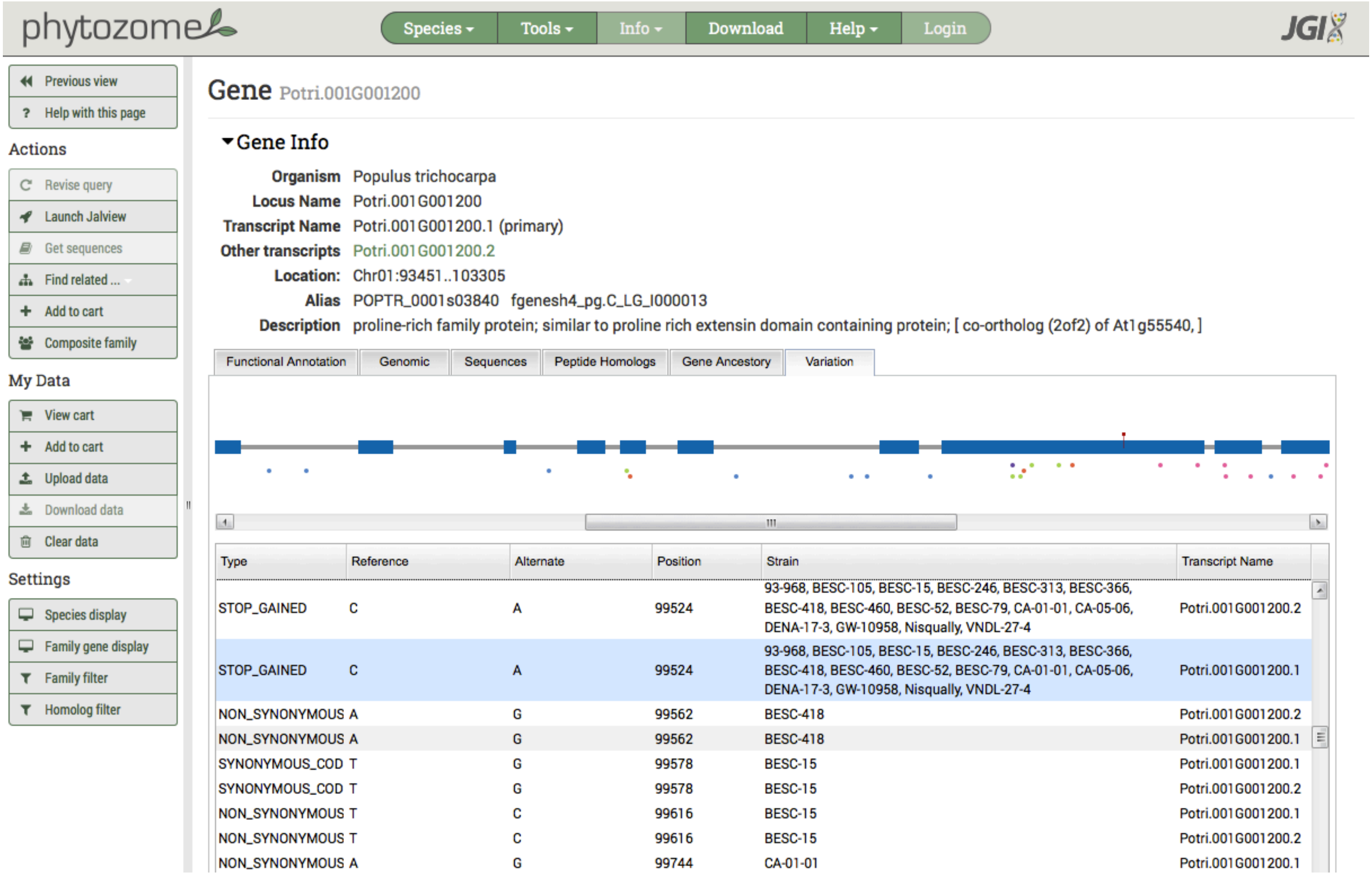
Genotypes:

Sample	Genotype	Total Depth
93-968	A/A	13
BESC-418	A/A	19
BESC-52	A/A	15
BESC-79	A/A	12
BESC-246	A/A	17
BESC-313	A/A	13
CA-05-06	A/A	10
BESC-460	A/A	11
GW-10958	A/A	14
Nisqually	C/A	19
DENA-17-3	A/A	29
CA-01-01	A/A	21
VNDL-27-4	A/A	18
BESC-15	A/A	16
BESC-105	A/A	19
BESC-366	A/A	7

SnpEff Variant Annotation, where available:

Effect	Effect Impact	Functional Class	Codon Change	Amino Acid Change	Gene Name	Transcript	Exon	Genotype
DOWNSTREAM	MODIFIER		1656		Potri.001G001300	Potri.001G001300.1		A
DOWNSTREAM	MODIFIER		1689		Potri.001G001300	Potri.001G001300.2		A
DOWNSTREAM	MODIFIER		1689		Potri.001G001300	Potri.001G001300.3		A
DOWNSTREAM	MODIFIER		2721		Potri.001G001300	Potri.001G001300.4		A
DOWNSTREAM	MODIFIER		4716		Potri.001G001100	Potri.001G001100.1		A
STOP_GAINED	HIGH	NONSENSE	tgC/tgA	C454*	Potri.001G001200	Potri.001G001200.2	8	A
STOP_GAINED	HIGH	NONSENSE	tgC/tgA	C454*	Potri.001G001200	Potri.001G001200.1	8	A
UPSTREAM	MODIFIER		4615		Potri.001G001400	Potri.001G001400.1		A
UPSTREAM	MODIFIER		4615		Potri.001G001400	Potri.001G001400.2		A

Diversity Data Mirrored in Phytozome Gene View



Gene Potri.001G001200

Gene Info

Organism Populus trichocarpa

Locus Name Potri.001G001200

Transcript Name Potri.001G001200.1 (primary)

Other transcripts Potri.001G001200.2

Location: Chr01:93451..103305

Alias POPTR_0001s03840 fgenesh4_pg.C_LG_I000013

Description proline-rich family protein; similar to proline rich extensin domain containing protein; [co-ortholog (2of2) of At1g55540,]

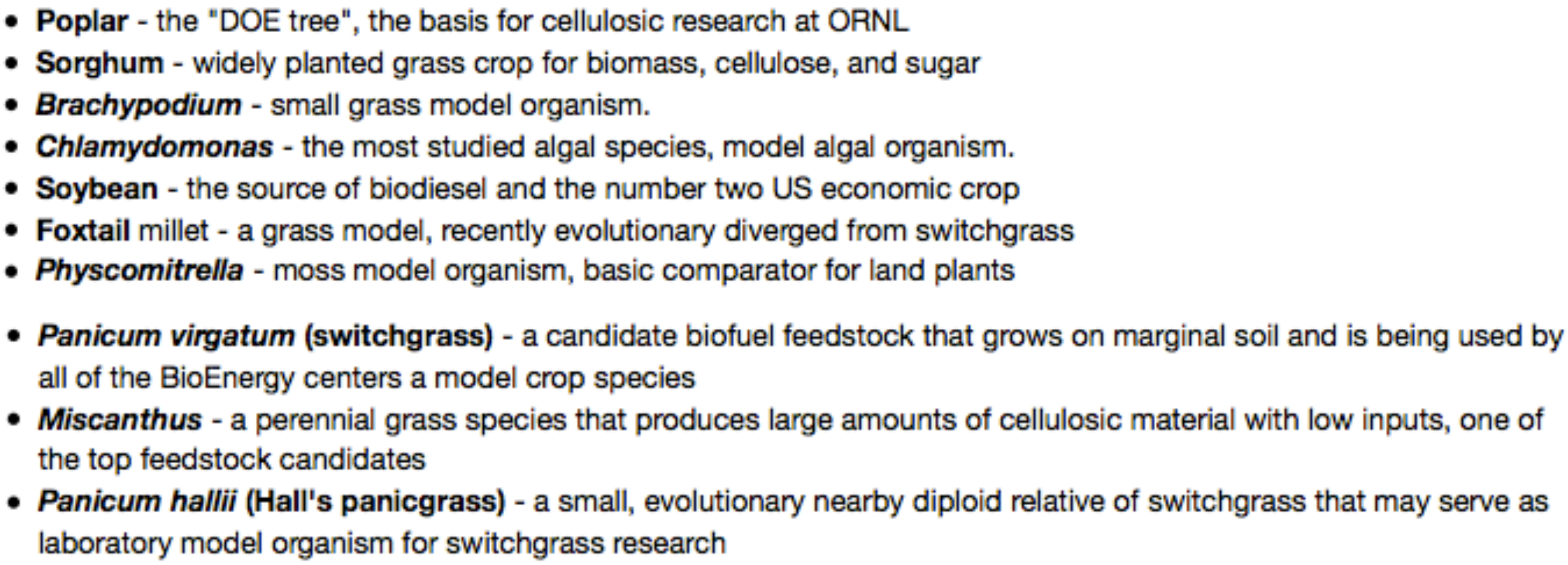
Type	Reference	Alternate	Position	Strain	Transcript Name
STOP_GAINED	C	A	99524	93-968, BESC-105, BESC-15, BESC-246, BESC-313, BESC-366, BESC-418, BESC-460, BESC-52, BESC-79, CA-01-01, CA-05-06, DENA-17-3, GW-10958, Nisqually, VNDL-27-4	Potri.001G001200.2
STOP_GAINED	C	A	99524	93-968, BESC-105, BESC-15, BESC-246, BESC-313, BESC-366, BESC-418, BESC-460, BESC-52, BESC-79, CA-01-01, CA-05-06, DENA-17-3, GW-10958, Nisqually, VNDL-27-4	Potri.001G001200.1
NON_SYNONYMOUS	A	G	99562	BESC-418	Potri.001G001200.2
NON_SYNONYMOUS	A	G	99562	BESC-418	Potri.001G001200.1
SYNONYMOUS_COD	T	G	99578	BESC-15	Potri.001G001200.1
SYNONYMOUS_COD	T	G	99578	BESC-15	Potri.001G001200.2
NON_SYNONYMOUS	T	C	99616	BESC-15	Potri.001G001200.1
NON_SYNONYMOUS	T	C	99616	BESC-15	Potri.001G001200.2
NON_SYNONYMOUS	A	G	99744	CA-01-01	Potri.001G001200.1

Overview

- Phytozome Introduction
- Track Data Conversion
- Name Indexing
- Diversity Data
- Expression Data
- Next Steps

Beyond Whole Genome Sequencing

Plant Gene Atlas



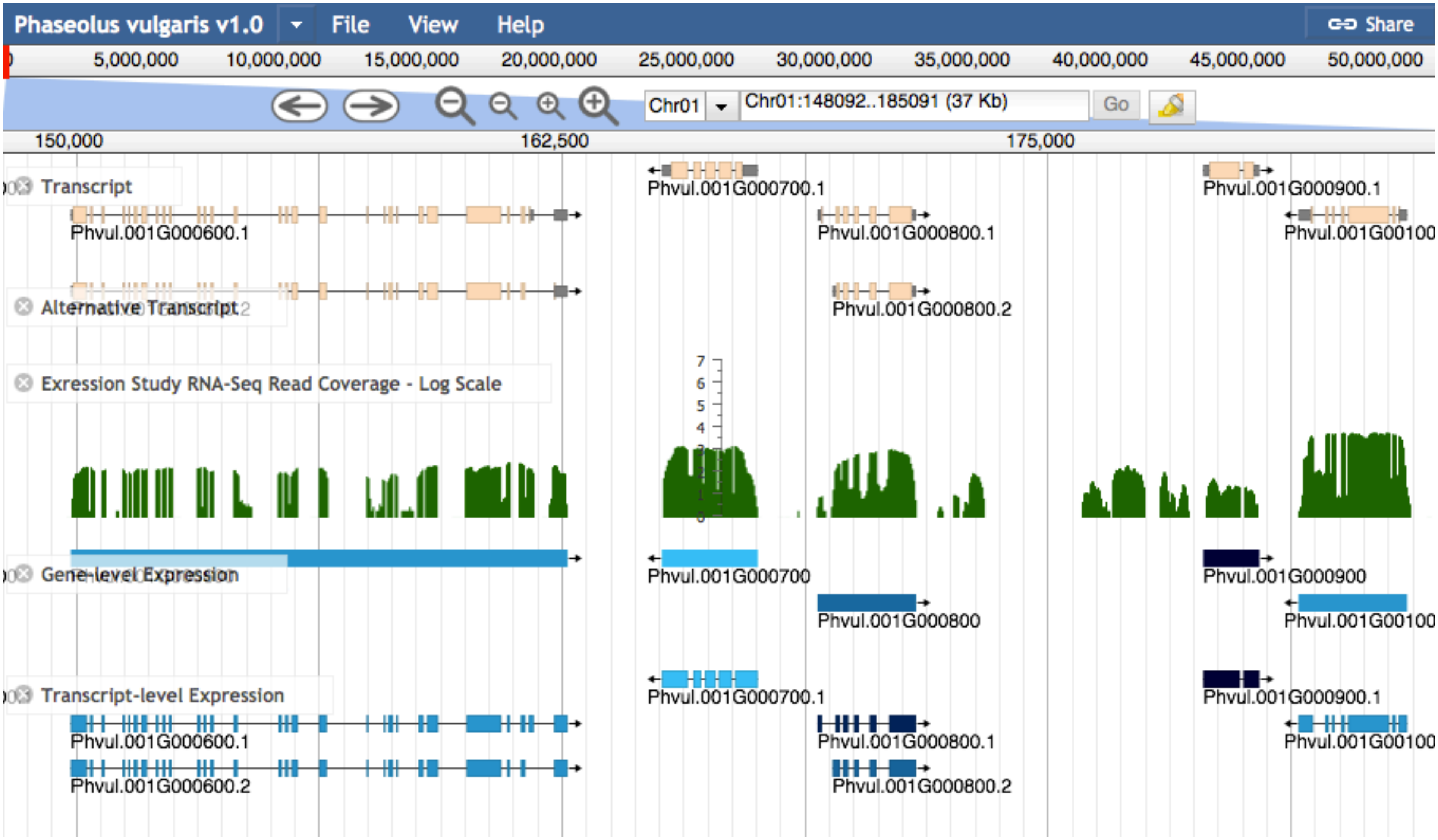
- **Poplar** - the "DOE tree", the basis for cellulosic research at ORNL
- **Sorghum** - widely planted grass crop for biomass, cellulose, and sugar
- ***Brachypodium*** - small grass model organism.
- ***Chlamydomonas*** - the most studied algal species, model algal organism.
- **Soybean** - the source of biodiesel and the number two US economic crop
- **Foxtail** millet - a grass model, recently evolutionary diverged from switchgrass
- ***Physcomitrella*** - moss model organism, basic comparator for land plants

- ***Panicum virgatum* (switchgrass)** - a candidate biofuel feedstock that grows on marginal soil and is being used by all of the BioEnergy centers a model crop species
- ***Miscanthus*** - a perennial grass species that produces large amounts of cellulosic material with low inputs, one of the top feedstock candidates
- ***Panicum hallii* (Hall's panicgrass)** - a small, evolutionary nearby diploid relative of switchgrass that may serve as laboratory model organism for switchgrass research

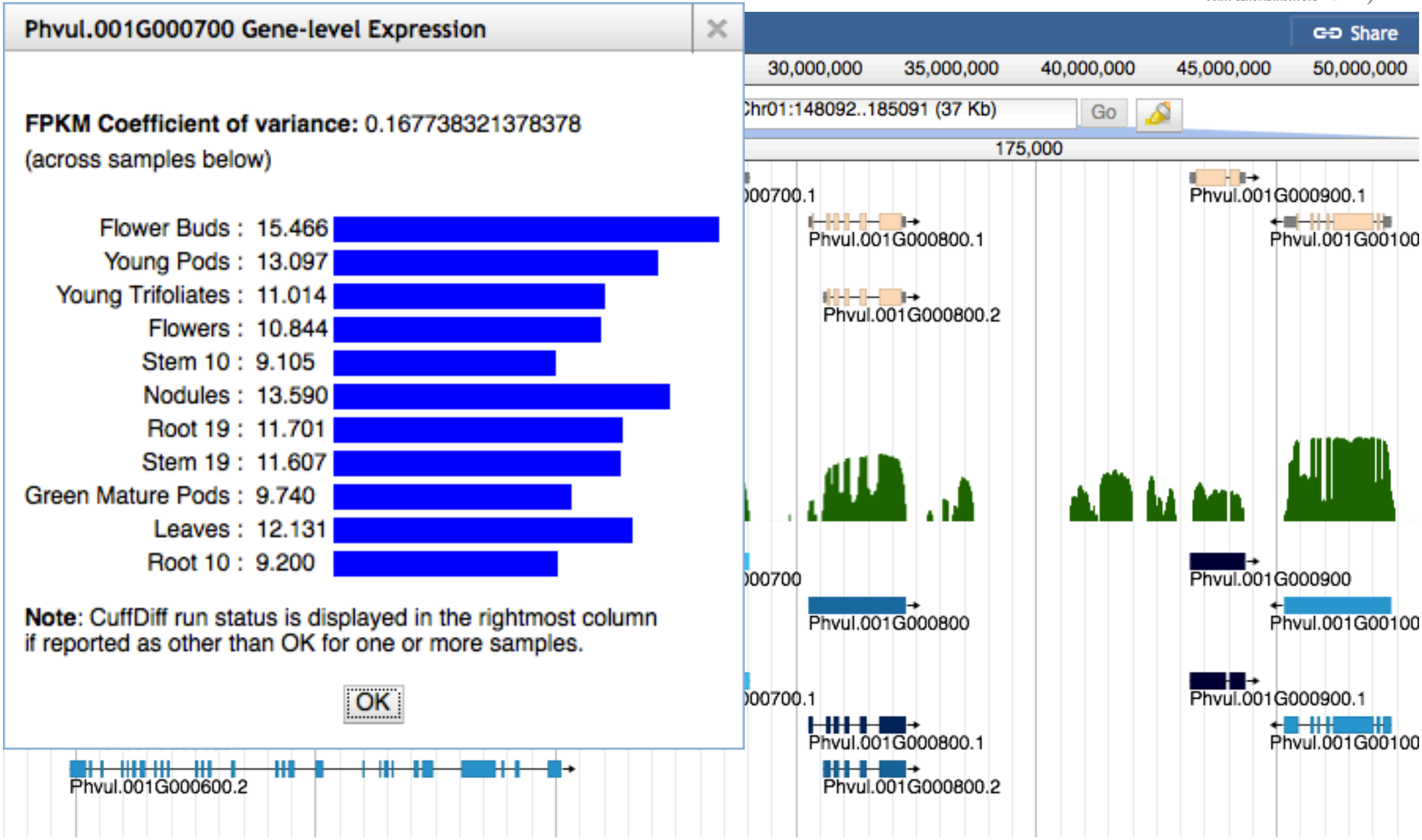
Transcriptome sequencing of varying growth conditions, different tissues and multiple developmental stages

- Read alignment, cuffdiff analysis to determine FPKM values indicating expression level variation

Phaseolus vulgaris example



Phaseolus vulgaris example



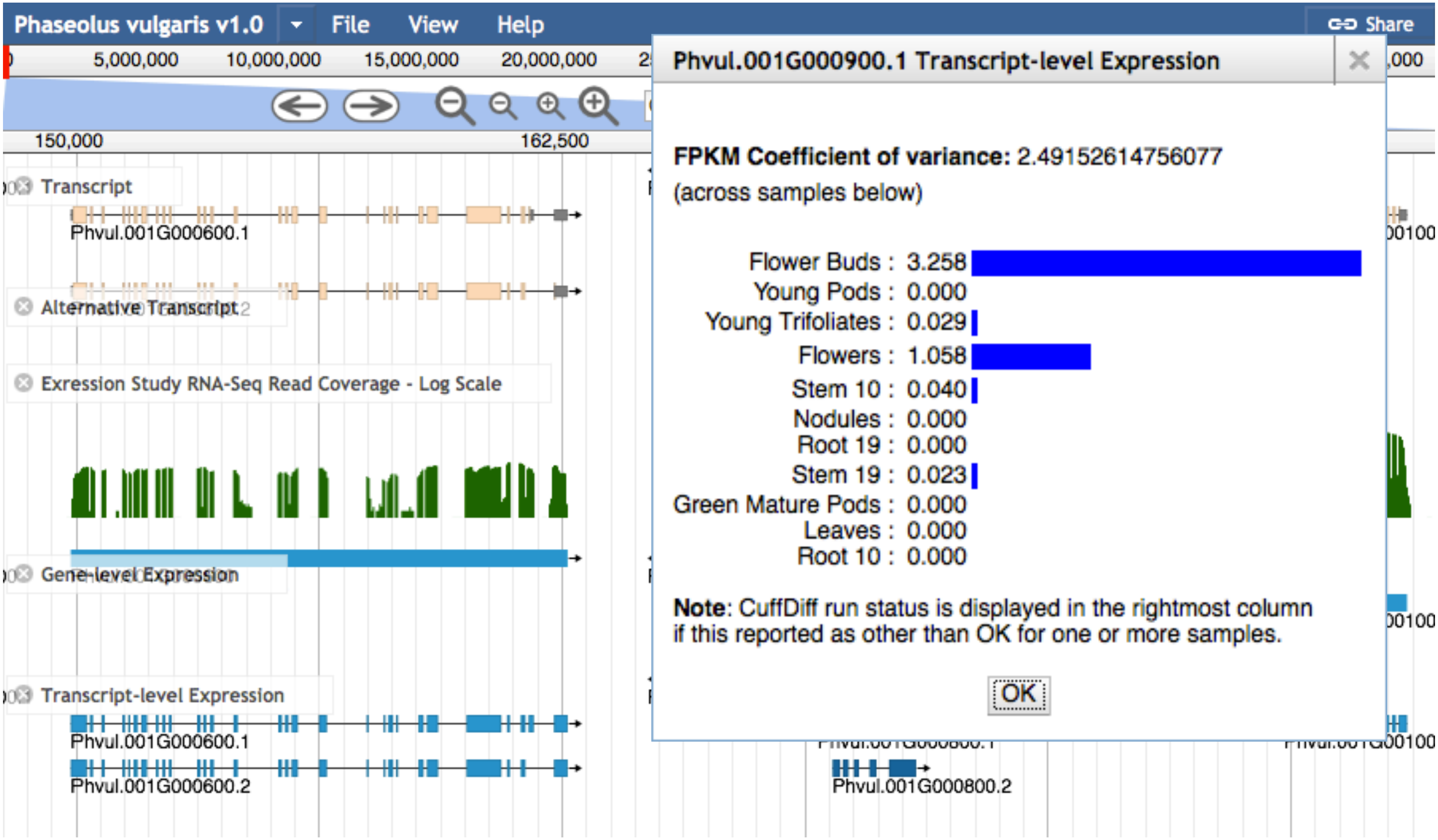
Phvul.001G000700 Gene-level Expression

FPKM Coefficient of variance: 0.167738321378378
(across samples below)

Flower Buds :	15.466
Young Pods :	13.097
Young Trifoliates :	11.014
Flowers :	10.844
Stem 10 :	9.105
Nodules :	13.590
Root 19 :	11.701
Stem 19 :	11.607
Green Mature Pods :	9.740
Leaves :	12.131
Root 10 :	9.200

Note: CuffDiff run status is displayed in the rightmost column if reported as other than OK for one or more samples.

Phaseolus vulgaris example



Phvul.001G000900.1 Transcript-level Expression

FPKM Coefficient of variance: 2.49152614756077 (across samples below)

Flower Buds :	3.258
Young Pods :	0.000
Young Trifoliates :	0.029
Flowers :	1.058
Stem 10 :	0.040
Nodules :	0.000
Root 19 :	0.000
Stem 19 :	0.023
Green Mature Pods :	0.000
Leaves :	0.000
Root 10 :	0.000

Note: CuffDiff run status is displayed in the rightmost column if this reported as other than OK for one or more samples.

OK

Summary

- **GFF3 conversion is pleasantly fast**

  Reconverted Primary and Alt. Transcript, BLAST, BLATX

  for all 45 genomes in 181 parallel cluster jobs in 2 hours

- **Name indexing continues to be the major data processing bottleneck**

  Much improved from just a few releases ago, however

  (run times varied 47 min to 2hr 50 min)

- **RNA-seq and SNP/indel call data analyses are quick to drop in to an existing JBrowse instance**

Acknowledgements

JGI

Dan Rokhsar
David Goodstein
Shengqiang Shu (Poplar)
Jeremy Phillips (GATK)
Ming Zhang (cuffdiff)

Rob Buels

JBrowse

GBrowse

GMOD

CHADO

biomart

InterMine

U.S. DEPARTMENT OF ENERGY | Office of Science