COMPARARATIVE GENOMICS WITH GMOD AND BIOPERL

Jason Stajich
Duke University

Fungal Comparative Genomics

- Problems
  - Many fungal genomes
  - No central place for annotations, interlinking homolog information
  - Want to visual gene structures and genome context
  - Need system for good database system for scripting genome questions

Getting the Data In

- GFF3 as the data transfer format
- Write GenBank -> GFF3 scripts
- Read in data from genome Centers (Broad, Sanger, WashU, JGI, SGD)
- Jason's Annotation Pipeline for Genome Annotatation

What I needed

- Database for storing and querying genome annotations
  - Bio::DB::GFF (BioPerl & Gbrowse)
- Visualization - Gbrowse
- Analyses
  - Ability to query for a gene's exon-intron structure and sequences
  - Are gene families clustered on chromosome?
  - Are functional classes of genes clustered on chromosome?

GBROWSE

- Visualization of annotation data
- Does not have to be for whole/finished genomes
  - Most projects are unfinished so many contigs (100s - 1000s)
  - BLAST interface with link to Gbrowse view allows user to start with query sequence and get to the genomic location

GBROWSE VIEW

Landmark or Region:

CHROMOSOME13:646044..6469 Search

Data Source

Cryptococcus neoformans var grubii serotype A, strain H99 (Duke 2004-10-30 assembly)

Reports & Analysis:

Annotate Restriction Sites Configure... Go

Scroll/Zoom: Show 936 bp

Overview

Overview of CHROMOSOME13

0k 100k 200k 300k 400k 500k 600k 700k

Genetic Markers

EcoV2.LG17 Xba6.LG17 Hind15.LG17

Sca2.LG17 Bam16.LG17

Bam7.LG17

Landmarks

Region

Details

646100 646200 646300 646400 646500 646600 646700 646800 646900

% gc

Exonerate est2genome

JEC21_CCAF044T.r1

SNAP predictions (JEC21 genomewise model)

CHROMOSOME13-snap.237

Twinscan predictions

TS.CHROMOSOME13.248

GLEAN models

GLEAN_01196

Probability 0.596427

GLEAN_gz_01176

Probability 0.999923

Exonerate protein2genome

cneo_JEC21_TIGR:CNN02110

anid_brd:AN7032.1

mgri_brd:MG04145.4.2

Pfam domains

Chitin_synth_2

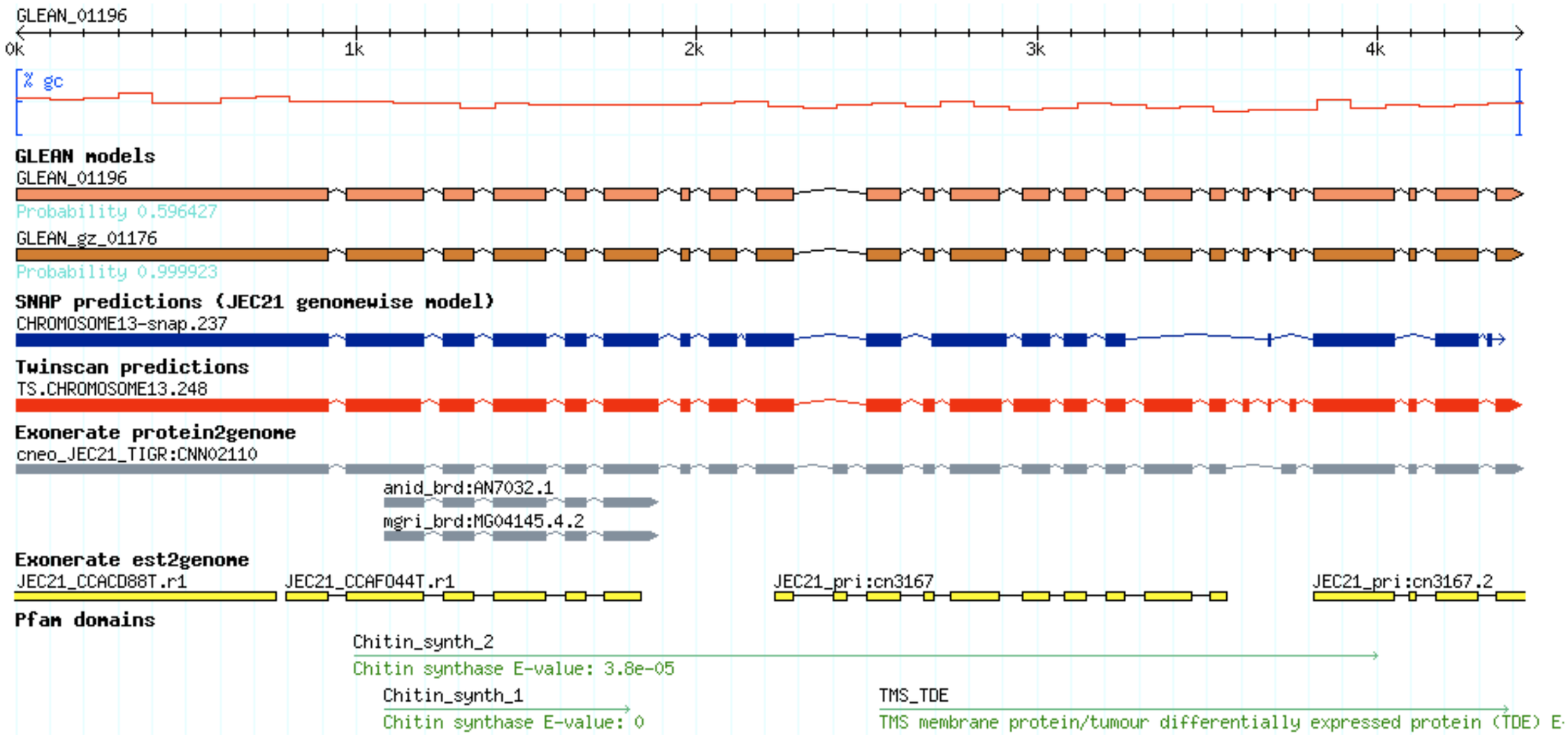
Chitin synthase E-value: 3.8e-05

Chitin_synth_1

Chitin synthase E-value: 0

Click Here

GENE PAGE (1)



Gbrowse Details

CDS Statistics

Locus Length: 4428 bp
CDS Length: 2952 bp
Exon count: 24 exons

Peptide Statistics

Protein Length: 984 residues
Molecular Weight: 108489.9

GENE PAGE (2)

Coding Sequence

```
>cneo_H99:GLEAN_01196
ATGCCAAACATATCACGCAAACCTCCTCCGCGcttctactctccttctcactccccttcgccttcacttt
ATGCTCCCATACAATCACCCCCGGCCCCCTCTTATGACTACCACGCCAACCCCAGGACGTTGAATCCATT
CTCCGACGCACGTGAAGTCGGTGGATATGCTCAACTCCAAGGGGAAGATCAAATGACTGGCGCACCTTTA
TACCAGCCTCCGTATGCTCCTCAACTACTCGTTGCTCAACCAActcctgtttcttcccgcctcccgtttt
tcGAAGCTGCGCTTGCCCGCGCGCGCGGTATCCAAACGCCCAGCTTACCAGAAGCCCCGACTCCAGCTTA
TGcccaacctttaccttcctacctcccgcctcccGACCCCAATCACCCTGATCTATCTGTTGGTTTGACT
CAGGCCAATACAGTTCGCTACGCCATCAATCCAAGATCTCAGTTGAAGGAAGGATCCCGTTCGCCTTCAC
CGTTCATGGATGACAGTTTCGTTTACAATGATGCTGCTCACCTCTATAACGTTGAGCCGGACGTCGAAAA
AGCTTTGCTCGGAAGTGGACTGGGTTATGAATCGGAGAAGCGTGTCGAATCTTCGATGGGCTTCAATGAC
AATGATGGCGACCTCTCAGTCCCTCAGTCGTTTGGTGGCCGACCTCCTTCATGGGAGCCGAGCGGCATAT
TGGATGAAAAAGGGGAAATGTCGACTACAAAGCATTTTGGGCCTGCACCTGCGGGTCGAGTCGGTCGGCG
AGCGCACAACGCTGCAGGGTACCGCAGGATCAAACAATCAGCGACCCTCGATGAGAATGGTTTCTTTGCT
ATCGAGATGAACATCCCCACCCGACTGGCGCAATTTCTACCCATCAAGGGAGTTGAAGAGCAAAAGACTA
CAAGGTATACTGCGATTACCACCGACCCAGATGATGTCCCAGCAGCTGGCTTCCGTCTTCGCCAGAACAT
GACTTCTCCGCCCCGACAGACTGAACTTTTTATCGTGATCACTATGTACAATGAGAACGCCGAGCTCTTT
TGTCGAACACTTTATGGTGTCATGAAGAATATAGCCCACCTATGTGGGCGTAAGAACTCAAGGGTCTGGG
GCAAGGATGGTTGGCAAAAGGTTGTCGTTTGCATTGTCGCGGACGGACGTAAGGCGGTTAACCCCCGCGT
CCTCGATTGTTTAGCAGCTCTTGGAGTTTACCAAGAAGGCGCAATGACGAACACAGTAAAGGATCGACCG
GTCACAGCGCATGTTTTCGAATACACGACCAGCTTTGCTCTTGACGGTGATTTACACTTCAAATATCCAG
ACAAAGGCATTGTCCCCTGCCAGATTATCTTCTGCATGAAAGAGAAAAATGCCAAAAAGATCAACTCCCA
TCGATGGTTTTTCAACGCCTTCGCGCCCTTGCTATCACCAAATGTCTGCATTCTTCTTGATGTGGGAACC
CAGCCAGCTCCGAAATCCATCTATCATCTTTGGAAAGCATTTGATGTCAATTCTAATGTTGGTGGTGCCT
GTGGAGAAATTGCGACCTTCAAGGGCAAAACTTGGAGGAGTTTATTGAACCCCCTTGTCGCGGCCCAAGC
CTTTGAGTACAAGATGTCCAACATCCTCGACAAACCTTTGGAGAGTCTCTTCGGATACTGCACTGTGTTG
CCTGGTGCCTTCTCGGCTTACAGGTGGATCGCTTTGCAAAACAATGGGGATGGGAGAACGGGACCTTTGG
CGAGTTATTTTGCTGGTGAACAGCTCAATACTGGAAAGGCAGACACATTCACTGGTAATATGGCCAAACC
CAAGGCCAACTGGGTGCTGAAATTCGTTAAGGCTGCTGTTGGAGAAACAGATTGCCCTGATACCATCCCA
GAGTTTATTGCTCAAAGAAGAAGATGGCTTAACGGTTCCTTCTTTGCAGCTGTCTATGCGTTGATGCACA
CGAACCAAATTTGGCGATCCGACCATTCGTTCGCGAGAAAGTCAGCCCTGATGTTGGAATCAGTGTACAA
CTTTCTGAACCTGATATTCTCGTGGTTCGCTTTGGCAAACTTTTACATTTTCTTTGTCATCCTTACGAGC
GCTTTGGAGGGCAGCGCTTTCAATGTCCCTCATATCGATGTGCTCAATACTATTGCACGATATGGTTACC
ttggtgctttggttggttgtttCATCTTCGCAATGGGAAACAGGCCACAAGGTTCGCCTTGGAAGTATAA
AGCAGCAATCTACTTTTTCGCCCTTTTGACTACCTATATGCTGGTCGCAGCAGTGCTTTGTACGGTACAG
GCAATCAAAAATATAAACAGCCCAATTTTTGCCAAGATGGTAGTATCACTCATATCAACCTATGGTATTT
ATGTGATTTCCAGTTTCTTGGCCCTTGACCCTTGGCACATCTTTACTTGCTTTATTCAATATGTTCTCTT
CTCACCTACTTATATCAATGTTCttaatgttTATGCCTATTCCAACCTTCACGACTTGTCATGGGGTACA
AAAGGCTCTGATGCAACCCAGGCGTCGGATTTGGGTGCTGTTTCCGGAGTGGGAAAGCACGTCGAAGTGG
AACTTGTAACTGCCCAGCAAGACATTGATATTGCCTATCAGGATGCTTTGGACAATATTAGATTAAGAGG
ATCAAAAGTTGACTCTGCTGAATCTGAGCCCAAAAAGGAGCAATCTGAACAAGCCCAGAAGGATACTTAT
GCCAACTTTCGTACCAATTTACTTTTGGTCTGGTCGCTGTCAAACGCCCTTCTCGCAAGTGTTATCCTTA
CAGGCAACAATTCTGGAGCGTTTGACGAGGGTTCCGGCAGTTCAAAAGCCACAATATACATGCTTGTGAT
```

Translation

```
>cneo_H99:GLEAN_01196
MPNISRKPPPRFYSPSHSPSPSLYAPIQSPPAPSYDYHANPRTLNPFSDAREVGGYAQLQGEDQMTGAPL
YQPPYAPQLLVAQPTPVSSRLPFFEAALARARGIQTPSLPEAPTPAYAQPLPSYLPPPDPNHPDLSVGLT
QANTVRYAINPRSQLKEGSRSPSPFMDDSFVYNDAAHLYNVEPDVEKALLGSGLGYESEKRVESSMGFND
NDGDLSVPQSFGGRPPSWEPSGILDEKGEMSTTKHFGPAPAGRVGRRAHNAAGYRRIKQSATLDENGFFA
IEMNIPTRLAQFLPIKGVEEQKTTRYTAITTDPDDVPAAGFRLRQNMTSPPRQTELFIVITMYNENAELF
CRTLYGVMKNIAHLCGRKNSRVWGKDGWQKVVVCIVADGRKAVNPRVLDCLAALGVYQEGAMTNTVKDRP
VTAHVFEYTTSFALDGDLHFKYPDKGIVPCQIIFCMKEKNAKKINSHRWFFNAFAPLLSPNVCILLDVGT
QPAPKSIYHLWKAFDVNSNVGGACGEIATFKGKTWRSLLNPLVAAQAFEYKMSNILDKPLESLFGYCTVL
PGAFSAYRWIALQNNGDGRTGPLASYFAGEQLNTGKADTFTGNMAKPKANWVLKFVKAAVGETDCPDTIP
EFIAQRRRWLNGSFFAAVYALMHTNQIWRSDHSFARKSALMLESVYNFLNLIFSWFALANFYIFFVILTS
ALEGSAFNVPHIDVLNTIARYGYLGALVGCFIFAMGNRPQGSPWKYKAAIYFFALLTTYMLVAAVLCTVQ
AIKNINSPIFAKMVVSLISTYGIYVISSFLALDPWHIFTCFIQYVLFSPTYINVLNVYAYSNLHDLSWGT
KGSDATQASDLGAVSGVGKHVEVELVTAQQDIDIAYQDALDNIRLRGSKVDSAESEPKKEQSEQAQKDTY
ANFRTNLLLVWSLSNALLASVILTGNNSGAFDEGSGSSKATIYMLVILIFVAGMSIFRFICSTLYLVISL
FTG*
```

GENE PAGE (3)

Intron sequences

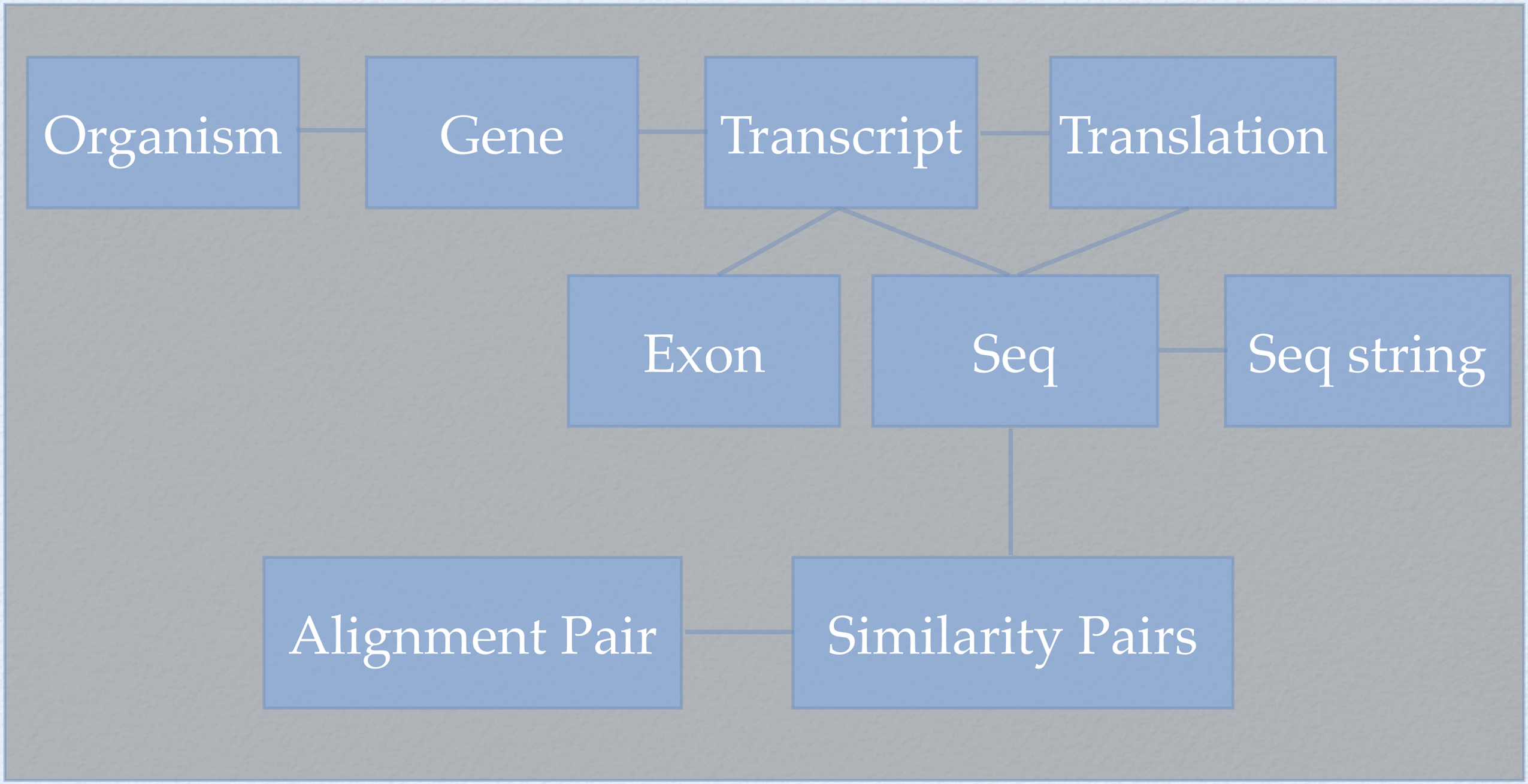
```
>cneo_H99:GLEAN_01196_intron1 CHROMOSOME13:645986..646042
GTAGGGCAGCGACTTTTGCAAGCTTGAGACCCCGTCTATTCGCTGACTCACACATAG
>cneo_H99:GLEAN_01196_intron2 CHROMOSOME13:646269..646324
GTGGGTGATCTTATATCTCCGCATATGTGTTTCAATATTGATGTCAAGGTATTCAG
>cneo_H99:GLEAN_01196_intron3 CHROMOSOME13:646413..646472
GTCAGGTTATCTTCAACAGTACAAAGCGCATTGCTGACATTTCGATACTTTGACAAATAG
>cneo_H99:GLEAN_01196_intron4 CHROMOSOME13:646625..646685
GTGAGCTTGGGGCTCCATCATGTATGCGTCATGTATTCGTAGCTGATTACTTCTTCTTTAG
>cneo_H99:GLEAN_01196_intron5 CHROMOSOME13:646743..646797
GTATGTATCACTCATTCGTTCGACCCAATGAACGCATCACTGACCATGTATGTAG
>cneo_H99:GLEAN_01196_intron6 CHROMOSOME13:646958..647023
GTAAGTCAGCAACCGCCGCACTAATATTCTACATGGTCAGCTAAACGCTGGTTTGTTGCCTACTAG
>cneo_H99:GLEAN_01196_intron7 CHROMOSOME13:647052..647107
GTGAGTACAACATGCAAATTTATTTGTGTCGAATCTGACGCTGATAATGACCATAG
>cneo_H99:GLEAN_01196_intron8 CHROMOSOME13:647187..647247
GTGAGTTGCAACAGCTGAGCATCTTTAGTTTCCAGGACTCACAGCAGACGGTGATATGCAG
>cneo_H99:GLEAN_01196_intron9 CHROMOSOME13:647355..647569
GTGAGTCAAAAATTGTAGAACTCAAGCGTTTTACTGACTGCCTCGTTCTTCATCAGGGTATGTTTCATAG
CGGCGCTTGATGTTTCTTTTTGATTATCTTCTAATCAGCTGCTATAGTACCTGGCAGAAGATAGAATCCT
GTGTTTCGAAATCGTGTAGGCACCACTTCACGGTAATAATACATGTCATTGTTGCTGATCCAACACGGCG
TATAG
>cneo_H99:GLEAN_01196_intron10 CHROMOSOME13:647672..647737
GTGAGTAATTCTTTTTATCggggaaaaaagaaaaaggggggTATCTGACATTCTTCGTCTTCTAG
>cneo_H99:GLEAN_01196_intron11 CHROMOSOME13:647766..647815
GTGAGCGCCGCACGGATTTGGCATTGAGTTTCACGCTTACATATCTTCAG
>cneo_H99:GLEAN_01196_intron12 CHROMOSOME13:647960..648030
GTGGGTACAGCCACTGTTGTATGTTTATACGGATACCCTAATAAGCaaaaaaaaaaaaaaaaaaaaaaCA
G
>cneo_H99:GLEAN_01196_intron13 CHROMOSOME13:648106..648151
GTACGCTGTTTTCATCCGTATAAGACATTAGCTCATTCGATGTTAG
>cneo_H99:GLEAN_01196_intron14 CHROMOSOME13:648214..648274
GTATGTGTTCATTTCTCTGGACAAGAGGGACAGCCAGCCGACGCTTTTCATCTTTCTTCAG
>cneo_H99:GLEAN_01196_intron15 CHROMOSOME13:648329..648389
GTGAGTCGTCACAAGTGGTGCTCAGGGGTCAAGGATAATCACTAAACGTTTTTTAACACAG
```

GENE PAGE (4)

Homologs from FASTA

Hit	Hit len	Bits	E-value	% sim	% id	% Query aligned
Cryptococcus neoformans - cneo_H99:GLEAN_01196	983	6060	0	100.0	100.0	100.0
Cryptococcus gattii - cneo_WM276:GLEAN_00366	984	5747	0	98.1	94.3	100.0
Cryptococcus gattii - cneo_R265:GLEAN_gz_05631	1009	3586	0	95.7	91.9	100.0
Cryptococcus neoformans - cneo_JEC21:CNN02110	996	3523	0	95.4	90.9	100.0
Phanerochaete chrysosporium - pchr:GLEAN_gz_10814	903	1775	1.6e-116	64.8	42.5	99.6
Uncinocarpus reesii - uree:GLEAN_05059	845	1590	1.7e-103	69.0	43.0	77.9
Phanerochaete chrysosporium - pchr:GLEAN_gz_12555	1004	1575	2.2e-102	74.2	52.9	83.9
Coprinus cinereus - ccin:GLEAN_gz2_06353	840	1514	3.7e-98	67.3	41.0	82.6
Coprinus cinereus - ccin:GLEAN_gz2_11986	1026	1495	9.7e-97	70.6	49.2	97.2
Botrytis cinerea - bcin:BC1G_11533	1168	1478	1.7e-95	65.4	42.0	99.5
Trichoderma reesei - tree:12480	955	1458	3.7e-94	68.4	45.6	88.2
Fusarium verticillioides - fver:GLEAN_09145	1180	1448	2.2e-93	66.2	41.8	98.9
Neurospora crassa - ncra:NCU05239.1	926	1436	1.3e-92	68.8	44.5	86.9
Histoplasma capsulatum - hcap_186R:GLEAN_05323	1149	1433	2.5e-92	58.4	38.2	99.3
Coprinus cinereus - ccin:GLEAN_gz2_06575	941	1430	3.4e-92	65.4	43.4	99.6
Uncinocarpus reesii - uree:GLEAN_08490	1210	1427	6.9e-92	65.8	41.8	99.0
Coccidioides immitis - cimm:anid_cimm_1.72-g26.1	1244	1423	1.3e-91	65.2	41.4	98.2
Phanerochaete chrysosporium - pchr:GLEAN_gz_04887	647	1420	1.3e-91	68.9	44.0	65.3

SIMILARITY DATABASE



OTHER TOOLS

- BLAST interface
  - Search your sequence and get marked up results with links to Gbrowse
- Retrieve Spliced CoDing sequences, translations, introns from Gbrowse

BLAST TOOL

Database and Program Op

Program tblastn Database

Overlay Hits over Genome Image

Enter sequence below (most sta 0 seqs at a time

>gi|1302185|emb|CAA96086.
MSDQNNRSRNEYHSNRKNEPSYE
TNMLYNGDDGNNNTINDNERDIY
VIQTTPELIHNGSQTMATPIERPFF
IPQYHDQPFGYNNGYHGLQAKDY
EYLHDDSRPVNDGKEELDSVKSGY
KESDIIVSNDNLTANRALKRSGTEI
VTCEPNQLAEKNFTVRQLKYLTPR
KKIVVCIISDGRSKINERSLALLSSL
GTVPIQLLFCLKEQNQKKINSHRW
IRTDLGKRFVKLLNPLVASQNFEYK
ENEGFHFFSSNMYLAEDRILCFEVV

Or load it from disk

Set subsequence: From

Clear sequence Search

The query sequence is filtered f
Filter Low complexity

Post Process with Smith-Wat

Expect 1 Matrix

Clear sequence Search

Powered by the WU-Blast Prog

- nt Archeascomycota
- nt Basidiomycota
- nt Cryptococcus
- ✓ nt Euascomycota
- nt Hemiascomycota
- nt Zygomycota
- nt ashbya_gossypii
- nt aspergillus_fumigatus
- nt aspergillus_nidulans
- nt aspergillus_terreus
- nt botrytis_cinerea
- nt candida_albicans
- nt candida_glabrata
- nt candida_guilliermondii
- nt candida_lusitaniae
- nt candida_tropicalis
- nt chaetomium_globosum
- nt coccidioides_immitis
- nt coprinus_cinereus
- nt cryptococcus_neoformans_H99
- nt cryptococcus_neoformans_JEC21
- nt cryptococcus_neoformans_R265
- nt cryptococcus_neoformans_WM276
- nt debaryomyces_hansenii
- nt fusarium_graminearum
- nt fusarium_verticillioides
- nt histoplasma_capsulatum_186R
- nt kluyveromyces_lactis
- nt kluyveromyces_waltii
- nt magnaporthe_grisea
- nt neurospora_crassa
- nt phanerochaete_chrysosporium
- nt pneumocystis_carnii
- nt podospora_anserina
- nt saccharomyces_bayanus
- nt saccharomyces_castellii
- nt saccharomyces_cerevisiae_rm11-1a_1
- nt saccharomyces_cerevisiae_s288c
- nt saccharomyces_cerevisiae_yjm789
- nt saccharomyces_kluyveri
- nt saccharomyces_kudriavzevii

RE-FORMATTED BLAST

TBLASTN Query of GI|1302185|EMB|CAA96086.1| against nt Euascomycota

TBLASTN 2.0MP-WashU [10-May-2005] [linux24-i686-ILP32F64 2005-05-10T21:16:37]

Reference: Gish, W. (1996-2000) http://blast.wustl.edu

Query= GI|1302185|EMB|CAA96086.1| CHS1 [SACCHAROMYCES CEREVISIAE]

(1,131 letters)

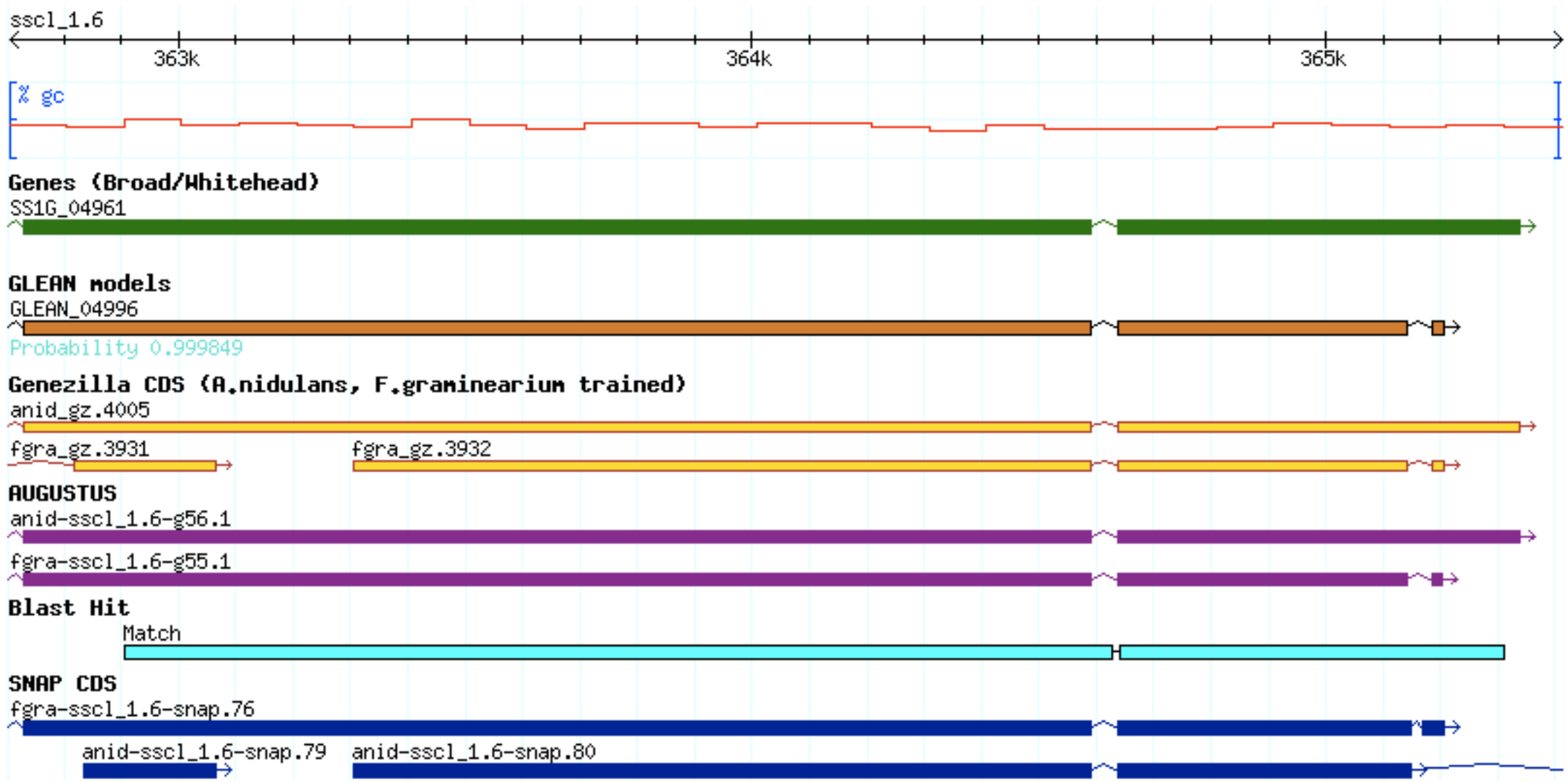
Database: uncinocarpus_reesii.2.nt; chaetomium_globosum.20041105.nt; coccidioides_immitis.20040311.nt; fusarium_gra
magnaporthe_grisea.20031031.nt; neurospora_crassa.20020212.nt; podospora_anserina.20040122.nt; aspergillus_fumiga
stagonospora_nodorum.20050205.nt; aspergillus_terreus.1.nt; fusarium_verticillioides.2.nt; sclerotinia_sclerotiorum.1.nt

12,142 sequences; 504,395,971 total letters

Sequences producing significant alignments:	Score (bits)	E value
sscl_1.6	1413	5.6e-143
snod_1.8	1381	1.4e-139
ncra:ncra_3.221	1366	5.2e-138
tree_50	1361	1.7e-137
fgra:fgra_1.425	1353	1.2e-136
fver_2.7	1347	5.5e-136
anid:anid_1.78	1312	2.8e-132
uree_2.4	1303	2.5e-131
hcap_186R:hcap-186R_17.30	1299	3.2e-131
cimm:cimm_1.106	1297	1.1e-130
afum:afum_57	1289	7.7e-130
ater_1.8	1286	1.6e-129

WITH LINKS

sscl_1.6 Link_group:2



Length = 2,040,374

Score = 502.5 bits (1413), Expect = 5.6e-143, P = 5.6e-143
Identities = 290/586 (49%), Positives = 386/586 (65%), Gaps = 23/586 (3%), Frame = +3
Links = (1)

```
Query: 314    KDDFSRDDEYDDLNTIDKLQFQANGVPASSSVSSIGSKESDIIVSNDNLTANRALKRSGT 373
              +D +  +D+ DD   I  LQ      P   + S    K  D    + + T   AL+R  T
Sbjct: 362907 QDPYGYNDD-DDHQPI--LQSHEPYGPDPHTASGAEYKGYDGAGHSPSSTPIPALRRYKT 363077
```

FETCH SEQUENCES

http://fungal.genome.duke.edu/cgi-bin/fg_get_sequence?type=CDS

Apple (4) ▾ News (1208) ▾ CSEM Cluster NCBI Web ▾ Science ▾ RSS ▾ Blogs ▾ OBI

NCBI Sequence Viewer v... | Retrieve sequence | Resources for Fungal Co...

Try YAL001C for S.cerevisiae

Fetch sequence by Identifier

Species Saccharomyces cerevisiae

Sequence Id YAL001C

Type CoDing Seuence (cds)

Reset submit

```
>scer_s288C_SGD:YAL001C scer-s288c_I:complement(join(151168
(YAL001C)
ATGGTACTGACGATTTATCCTGACGAACTCGTACAAATAGTGTCTGATAAAATTGCTTC
AATAAGGGAAAAATCACTTTGAATCAGCTGTGGGATATATCTGGTAAATATTTTGATTT
TCTGATAAAAAAGTTAAACAGTTCGTGCTTTCATGCGTGATATTGAAAAAGGACATTGA
GTGTATTGTGATGGTGCTATAACAACTAAAAATGTGACTGATATTATAGGCGACGCTAA
CATTCATACTCGGTTGGGATTACTGAGGACAGCCTATGGACATTATTAACGGGATACAC
AAAAAGGAGTCAACTATTGGAAATTCTGCATTTGAACTACTTCTCGAAGTTGCCAAATC
GGAGAAAAAGGGATCAATACTATGGATTTGGCGCAGGTAACTGGGCAAGATCCTAGAAC
GTGACTGGACGTATCAAGAAAATAAACCACCTGTTAACAAGTTCACAACTGATTTATAA
GGACACGTCGTGAAGCAATTGAAGCTAAAAAAATTCAGCCATGACGGGGTGGATAGTAA
CCCTATATTAATATTAGGGATCATTTAGCAACAATAGTTGAGGTGGTAAAACGATCAAA
AATGGTATTCGCCAGATAATTGATTTAAAGCGTGAATTGAAATTTGACAAAGAGAAAAC
CTTTCTAAAGCTTTTATTGCAGCTATTGCATGGTTAGATGAAAAGGAGTACTTAAAGAA
GTGCTTGTAGTATCACCCAAGAATCCTGCCATTAAAATCAGATGTGTAAAATACGTGAA
GATATTCCAGACTCTAAAGGCTCGCCTTCATTTGAGTATGATAGCAATAGCGCGGATGA
GATTCTGTATCAGATAGCAAGGCAGCTTTCGAAGATGAAGACTTAGTCGAAGGTTTAGA
AATTTCAATGCGACTGATTTATTACAAAATCAAGGCCTTGTTATGGAAGAGAAAGAGGA
GCTGTAAAGAATGAAGTTCTTCTTAATCGATTTTATCCACTTCAAAATCAGACTTATGA
ATTGCAGATAAGTCTGGCCTTAAAGGAATTTCAACTATGGATGTTGTAAATCGAATTAC
GGAAAAGAATTTCAGCGAGCTTTTACCAAATCAAGCGAATATTATTTAGAAAGTGTGGA
AAGCAAAAAGAAAATACAGGGGGGTATAGGCTTTTTCGCATATACGATTTTGAGGGAAA
AAGAAGTTTTTTAGGCTGTTCACAGCTCAGAACTTTCAAAAGTTAACAAATGCGGAAGA
GAAATATCCGTTCCAAAAGGGTTTGATGAGCTAGGCAAATCTCGTACCGATTTGAAAAC
CTCAACGAGGATAATTTCGTCGCACTCAACAACACTGTTAGATTTACAACGGACAGCGA
GGACAGGATATATTCTTCTGGCACGGTGAATTAAAAATTCCCCCAAACTCAAAAAAAAC
CCGAATAAAAACAAACGGAAGAGGCAGGTTAAAAACAGTACTAATGCTTCTGTTGCAGC
AACATTTCGAATCCCAAAAGGATTAAGCTAGAGCAGCATGTCAGCACTGCACAGGAGCC
AAATCTGCTGAAGATAGTCCAAGTTCAAACGGAGGCACTGTTGTCAAAGGCAAGGTGGT
AACTTCGGCGGCTTTTCTGCCCGCTCTTTGCGTTCACTACAGAGACAGAGAGCCATTTT
AAAGTTATGAATACGATTGGTGGGGTAGCATACCTGAGAGAACAATTTTACGAAAGCGT
TCTAAATATATGGGCTCCACAACGACATTAGATAAAAAGACTGTCCGTGGTGATGTTGA
TTGATGGTAGAAAGCGAAAAATTAGGAGCCAGAACAGAGCCTGTATCAGGAAGAAAAAT
ATTTTTTTGCCCACTGTTGGAGAGGACGCTATCCAAAGGTACATCCTGAAAGAAAAAGA
AGTAAAAAAGCAACCTTTACTGATGTTATACATGATACGGAAATATACTTCTTTGACCA
ACGGAAAAAAATAGGTTTCACAGAGGAAAGAAATCAGTTGAAAGAATTCGTAAGTTTCA
```

What's Missing

- Homolog/Ortholog/Paralog capturing
  - Pairwise focused summary stastistics
  - Multiway ortholog summaries
  - Ensembl Compara --> GMOD Compara?
  - Linking to gene trees

Tools Used

- Perl & BioPerl (www.bioperl.org)
- Gbrowse (www.gmod.org/ggb)
- Mysql database