H99 Genome Update

Jason Stajich
Duke University

A Summary of Progress

- Sequencing, Assembly, and Finishing
- Automated Gene Annotation
- Comparative analyses

H99 Genome

- 11X Genome Coverage
- BAC End sequences
- FPC map
- 1st Broad Assembly (May-2003)
  - 19.2 Mb, 341 contigs
- Duke Oct-2004 assembly
  - 18.9 Mb, 14 chromosomes, Mito

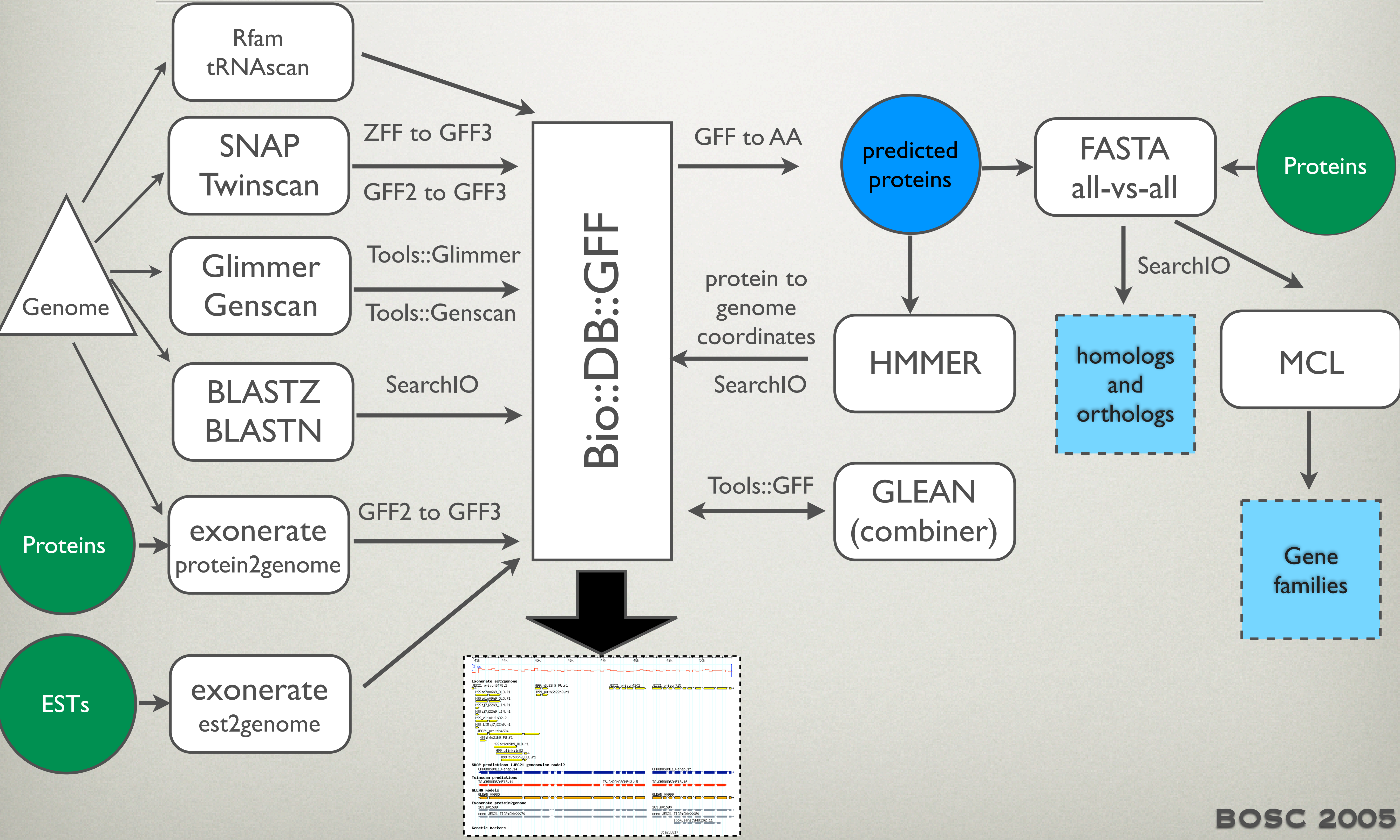
Finishing @Duke

- Andria Allen, Fred Dietrich
- Gap closure
- Re-Assembly

H99 Genome Annotation

- Protein coding gene predictions
  - *Ab initio*
    - SNAP (Korf, 2004) - trained on JEC21 annotations
    - Twinscan (Flicek et al, 2003; Tenney et al, 2004)
  - JEC21 proteins mapped
  - Genewise, exonerate
  - Combined predictions (GLEAN)
- RNA gene predictions (Rfam)

Genome Annotation Pipeline



Genome

Rfam
tRNAscan

SNAP
Twinscan

ZFF to GFF3

GFF2 to GFF3

Glimmer
Genscan

Tools::Glimmer

Tools::Genscan

BLASTZ
BLASTN

SearchIO

Proteins

exonerate
protein2genome

GFF2 to GFF3

ESTs

exonerate
est2genome

Bio::DB::GFF

GFF to AA

predicted proteins

protein to genome coordinates

SearchIO

HMMER

Tools::GFF

GLEAN (combiner)

FASTA
all-vs-all

Proteins

SearchIO

homologs and orthologs

MCL

Gene families

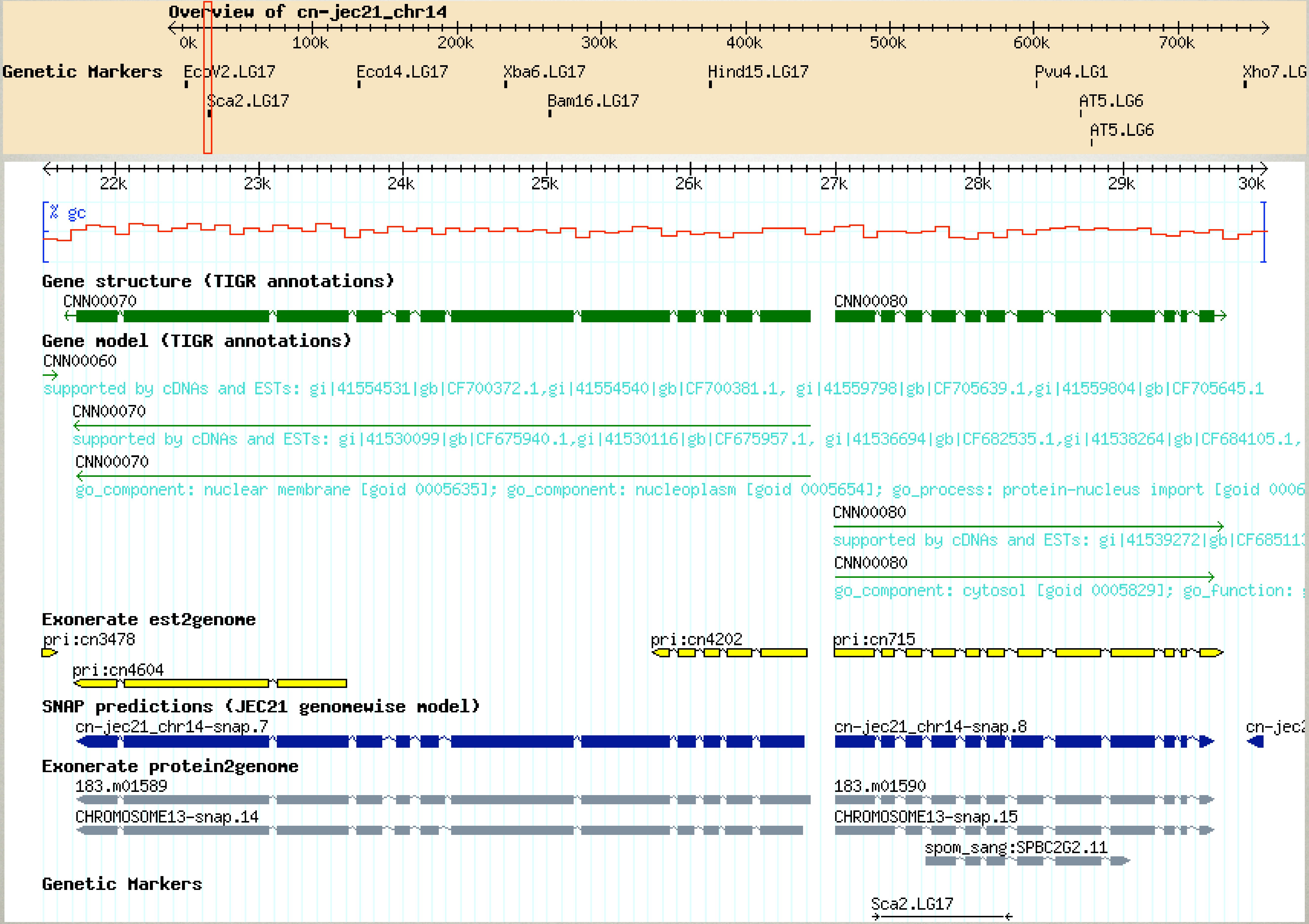
Gene Summary

- 7066 genes from SNAP
- 7357 genes from Twinscan
- 28k proteins mapped, ~60k ESTs mapped

- 6626 genes from GLEAN (consensus)

Website for Browsing Fungal Genomes

- http://fungal.genome.duke.edu
- Gbrowse view of many fungal genomes
- Annotations for H99, JEC21, WM276, R265
- BLAST against annotations and genomes
  - See Hits in GBrowse context



Overview of cn-jec21_chr14

Genetic Markers

Gene structure (TIGR annotations)

Gene model (TIGR annotations)

CNN00060

supported by cDNAs and ESTs: gi|41554531|gb|CF700372.1,gi|41554540|gb|CF700381.1, gi|41559798|gb|CF705639.1,gi|41559804|gb|CF705645.1

CNN00070

supported by cDNAs and ESTs: gi|41530099|gb|CF675940.1,gi|41530116|gb|CF675957.1, gi|41536694|gb|CF682535.1,gi|41538264|gb|CF684105.1,

CNN00070

go_component: nuclear membrane [goid 0005635]; go_component: nucleoplasm [goid 0005654]; go_process: protein-nucleus import [goid 0006

CNN00080

supported by cDNAs and ESTs: gi|41539272|gb|CF68511

CNN00080

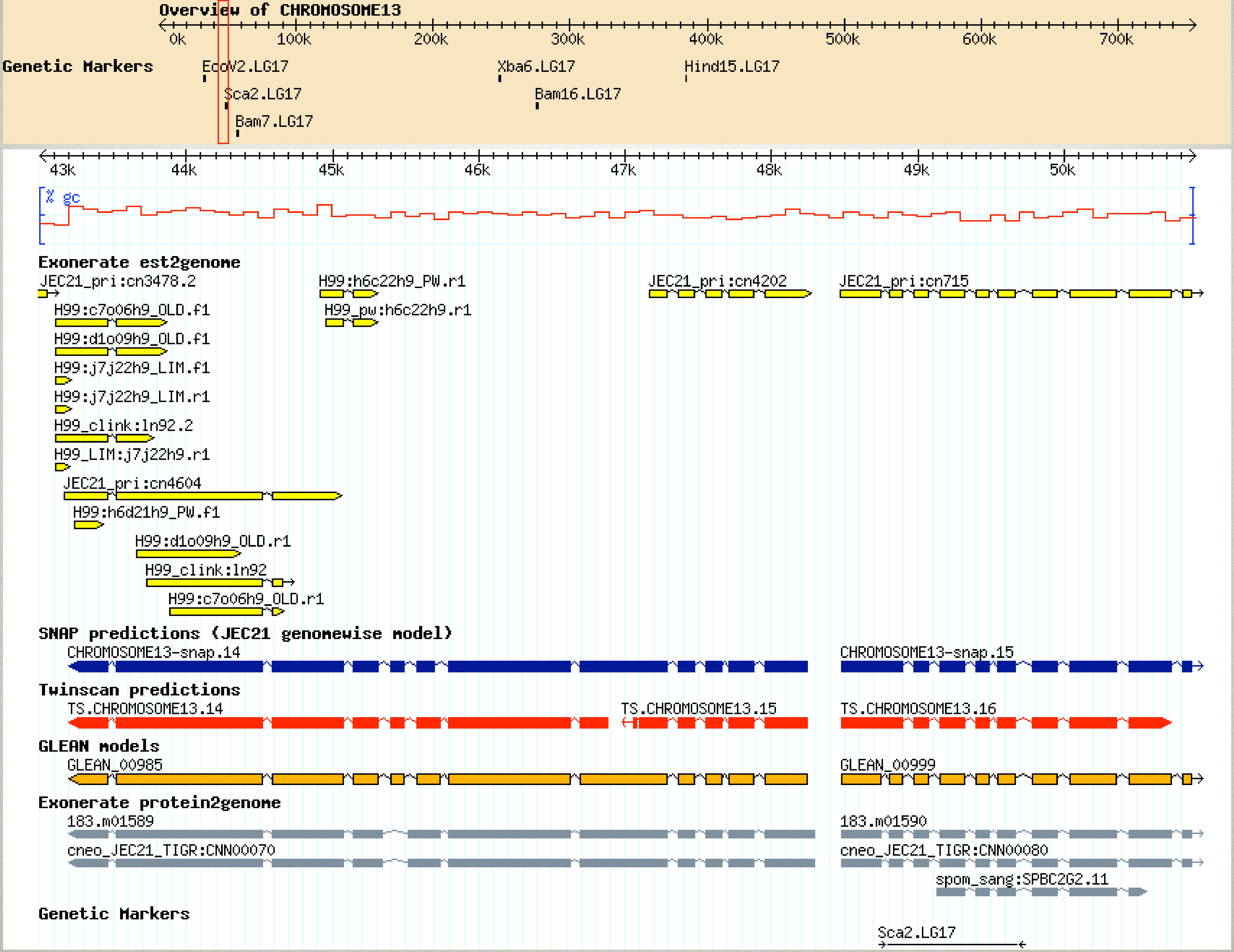
go_component: cytosol [goid 0005829]; go_function:

Exonerate est2genome

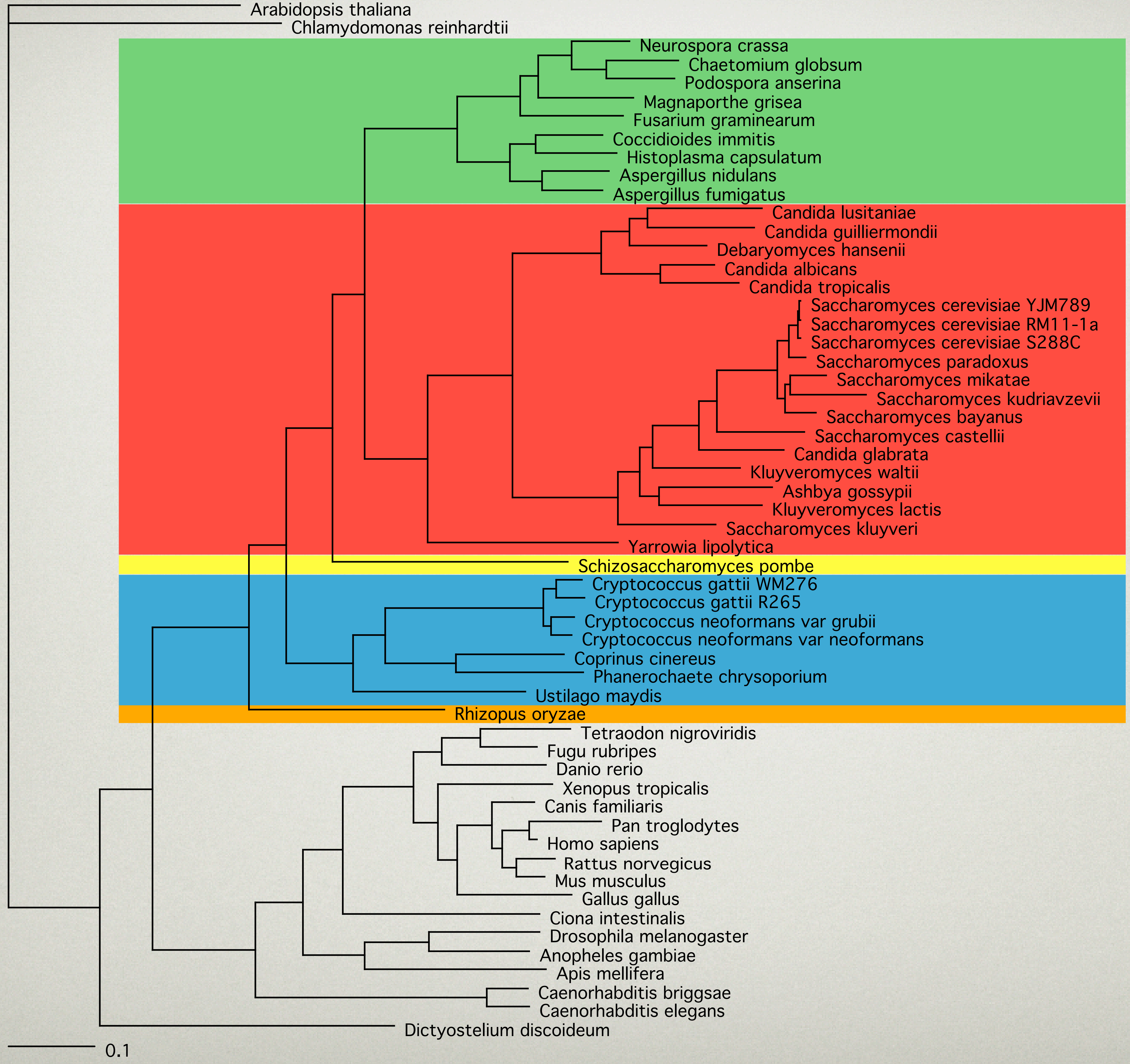
SNAP predictions (JEC21 genomewise model)

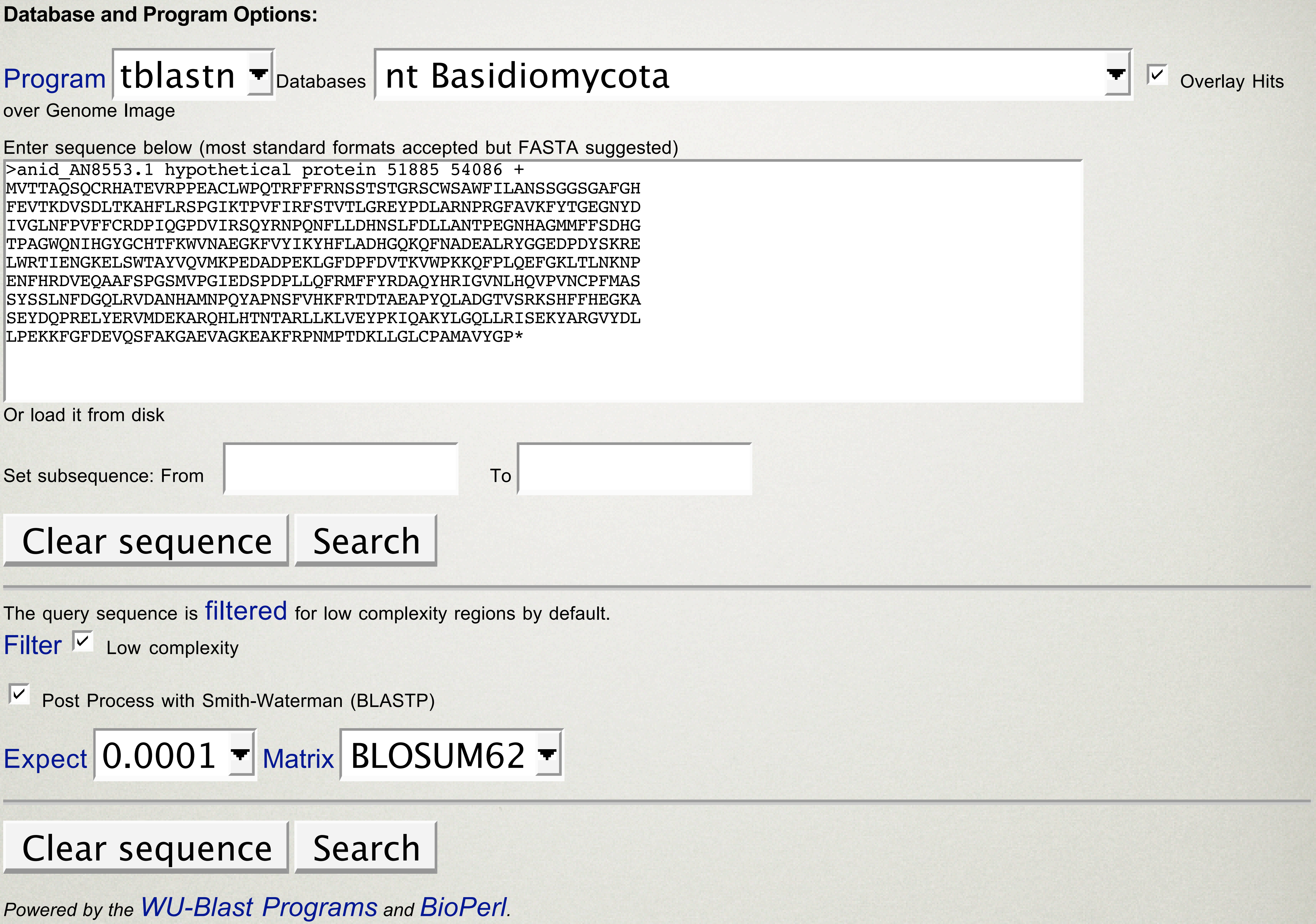
Exonerate protein2genome

Genetic Markers


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
43k
44k
45k
46k
47k
48k
49k
50k
% gc
Exonerate est2genome
JEC21_pri:cn3478.2
H99:h6c22h9_PW.r1
JEC21_pri:cn4202
JEC21_pri:cn715
H99:c7o06h9_OLD.f1
H99_pw:h6c22h9.r1
H99:d1o09h9_OLD.f1
H99:j7j22h9_LIM.f1
H99:j7j22h9_LIM.r1
H99_clink:ln92.2
H99_LIM:j7j22h9.r1
JEC21_pri:cn4604
H99:h6d21h9_PW.f1
H99:d1o09h9_OLD.r1
H99_clink:ln92
H99:c7o06h9_OLD.r1
SNAP predictions (JEC21 genomewise model)
CHROMOSOME13-snap.14
CHROMOSOME13-snap.15
Twinscan predictions
TS.CHROMOSOME13.14
TS.CHROMOSOME13.15
TS.CHROMOSOME13.16
GLEAN models
GLEAN_00985
GLEAN_00999
Exonerate protein2genome
183.m01589
183.m01590
cneo_JEC21_TIGR:CNN00070
cneo_JEC21_TIGR:CNN00080
spom_sang:SPBC2G2.11
Genetic Markers
Sca2.LG17

Comparative Analyses


Arabidopsis thaliana
Chlamydomonas reinhardtii
Neurospora crassa
Chaetomium globsum
Podospora anserina
Magnaporthe grisea
Fusarium graminearum
Coccidioides immitis
Histoplasma capsulatum
Aspergillus nidulans
Aspergillus fumigatus
Candida lusitaniae
Candida guilliermondii
Debaryomyces hansenii
Candida albicans
Candida tropicalis
Saccharomyces cerevisiae YJM789
Saccharomyces cerevisiae RM11-1a
Saccharomyces cerevisiae S288C
Saccharomyces paradoxus
Saccharomyces mikatae
Saccharomyces kudriavzevii
Saccharomyces bayanus
Saccharomyces castellii
Candida glabrata
Kluyveromyces waltii
Ashbya gossypii
Kluyveromyces lactis
Saccharomyces kluyveri
Yarrowia lipolytica
Schizosaccharomyces pombe
Cryptococcus gattii WM276
Cryptococcus gattii R265
Cryptococcus neoformans var grubii
Cryptococcus neoformans var neoformans
Coprinus cinereus
Phanerochaete chrysoporium
Ustilago maydis
Rhizopus oryzae
Tetraodon nigroviridis
Fugu rubripes
Danio rerio
Xenopus tropicalis
Canis familiaris
Pan troglodytes
Homo sapiens
Rattus norvegicus
Mus musculus
Gallus gallus
Ciona intestinalis
Drosophila melanogaster
Anopheles gambiae
Apis mellifera
Caenorhabditis briggsae
Caenorhabditis elegans
Dictyostelium discoideum
0.1



Database and Program Options:

Program: tblastn Databases: nt Basidiomycota ☑ Overlay Hits over Genome Image

Enter sequence below (most standard formats accepted but FASTA suggested)

```
>anid_AN8553.1 hypothetical protein 51885 54086 +
MVTTAQSQCRHATEVRPPEACLWPQTRFFFRNSSTSTGRSCWSAWFILANSSGGSGAFGH
FEVTKDVSDLTKAHFLRSPGIKTPVFIRFSTVTLGREYPDLARNPRGFAVKFYTGEGNYD
IVGLNFPVFFCRDPIQGPDVIRSQYRNPQNFLLDHNSLFDLLANTPEGNHAGMMFFSDHG
TPAGWQNIHGYGCHTFKWVNAEGKFVYIKYHFLADHGQKQFNADEALRYGGEDPDYSKRE
LWRTIENGKELSWTAYVQVMKPEDADPEKLGFDPFDVTKVWPKKQFPLQEFGKLTLNKNP
ENFHRDVEQAAFSPGSMVPGIEDSPDPLLQFRMFFYRDAQYHRIGVNLHQVPVNCPFMAS
SYSSLNFDGQLRVDANHAMNPQYAPNSFVHKFRTDTAEAPYQLADGTVSRKSHFFHEGKA
SEYDQPRELYERVMDEKARQHLHTNTARLLKLVEYPKIQAKYLGQLLRISEKYARGVYDL
LPEKKFGFDEVQSFAKGAEVAGKEAKFRPNMPTDKLLGLCPAMAVYGP*
```

Or load it from disk

Set subsequence: From To

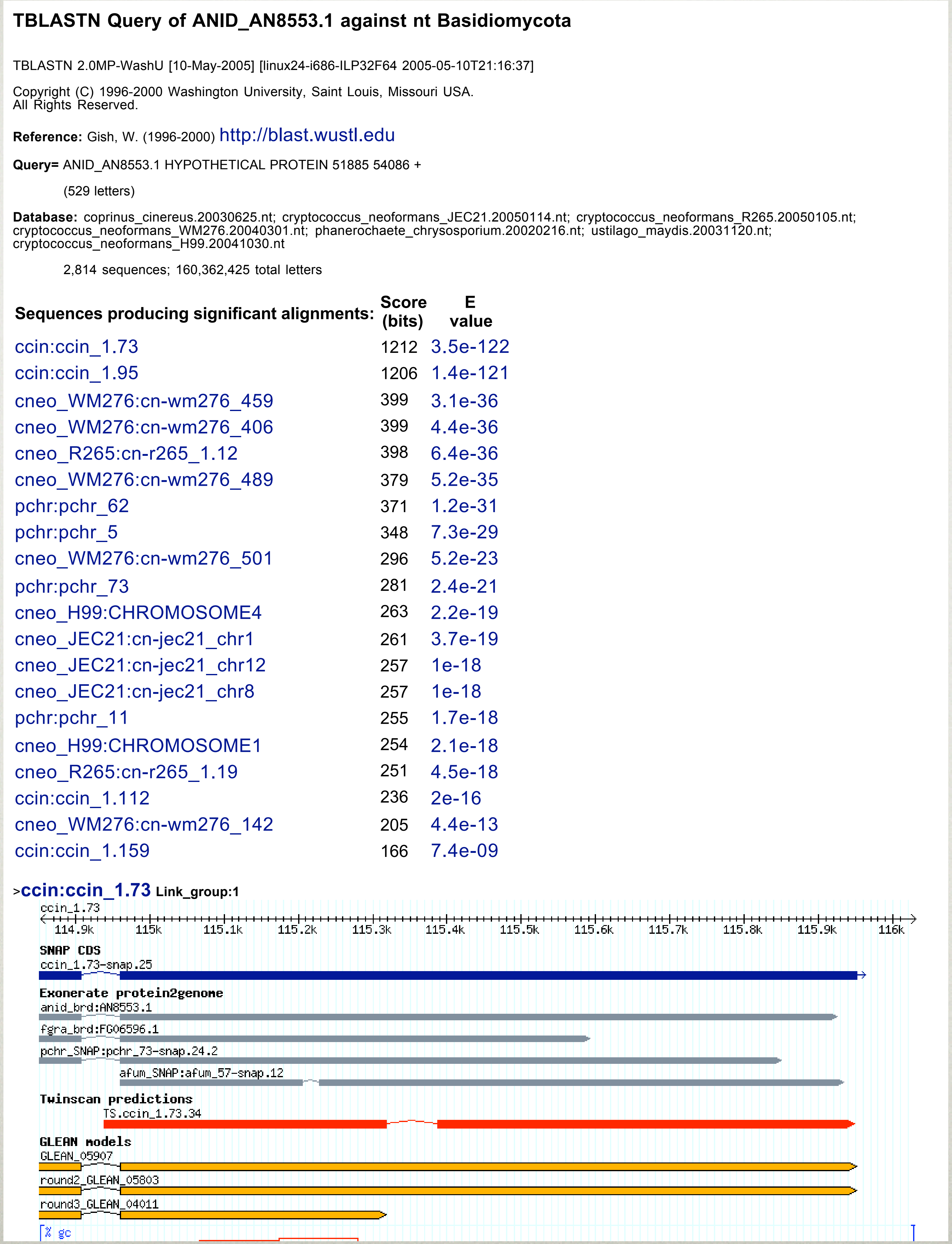
The query sequence is filtered for low complexity regions by default.

Filter ☑ Low complexity

☑ Post Process with Smith-Waterman (BLASTP)

Expect: 0.0001 Matrix: BLOSUM62

Powered by the WU-Blast Programs and BioPerl.



TBLASTN Query of ANID_AN8553.1 against nt Basidiomycota

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

Copyright (C) 1996-2000 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

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

Query= ANID_AN8553.1 HYPOTHETICAL PROTEIN 51885 54086 +

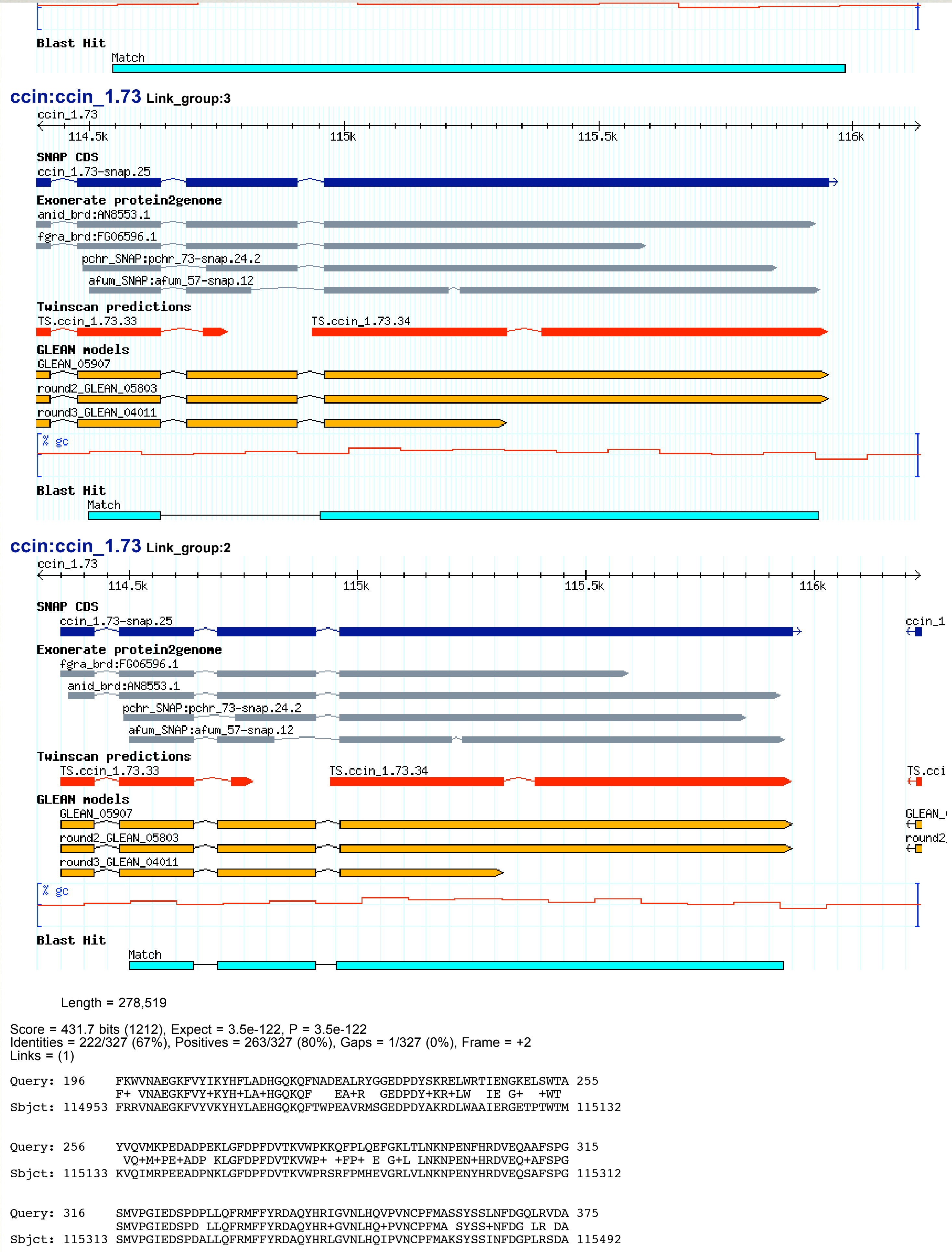
(529 letters)

Database: coprinus_cinereus.20030625.nt; cryptococcus_neoformans_JEC21.20050114.nt; cryptococcus_neoformans_R265.20050105.nt; cryptococcus_neoformans_WM276.20040301.nt; phanerochaete_chrysosporium.20020216.nt; ustilago_maydis.20031120.nt; cryptococcus_neoformans_H99.20041030.nt

2,814 sequences; 160,362,425 total letters

Sequences producing significant alignments:	Score (bits)	E value
ccin:ccin_1.73	1212	3.5e-122
ccin:ccin_1.95	1206	1.4e-121
cneo_WM276:cn-wm276_459	399	3.1e-36
cneo_WM276:cn-wm276_406	399	4.4e-36
cneo_R265:cn-r265_1.12	398	6.4e-36
cneo_WM276:cn-wm276_489	379	5.2e-35
pchr:pchr_62	371	1.2e-31
pchr:pchr_5	348	7.3e-29
cneo_WM276:cn-wm276_501	296	5.2e-23
pchr:pchr_73	281	2.4e-21
cneo_H99:CHROMOSOME4	263	2.2e-19
cneo_JEC21:cn-jec21_chr1	261	3.7e-19
cneo_JEC21:cn-jec21_chr12	257	1e-18
cneo_JEC21:cn-jec21_chr8	257	1e-18
pchr:pchr_11	255	1.7e-18
cneo_H99:CHROMOSOME1	254	2.1e-18
cneo_R265:cn-r265_1.19	251	4.5e-18
ccin:ccin_1.112	236	2e-16
cneo_WM276:cn-wm276_142	205	4.4e-13
ccin:ccin_1.159	166	7.4e-09

>**ccin:ccin_1.73** **Link_group:1**



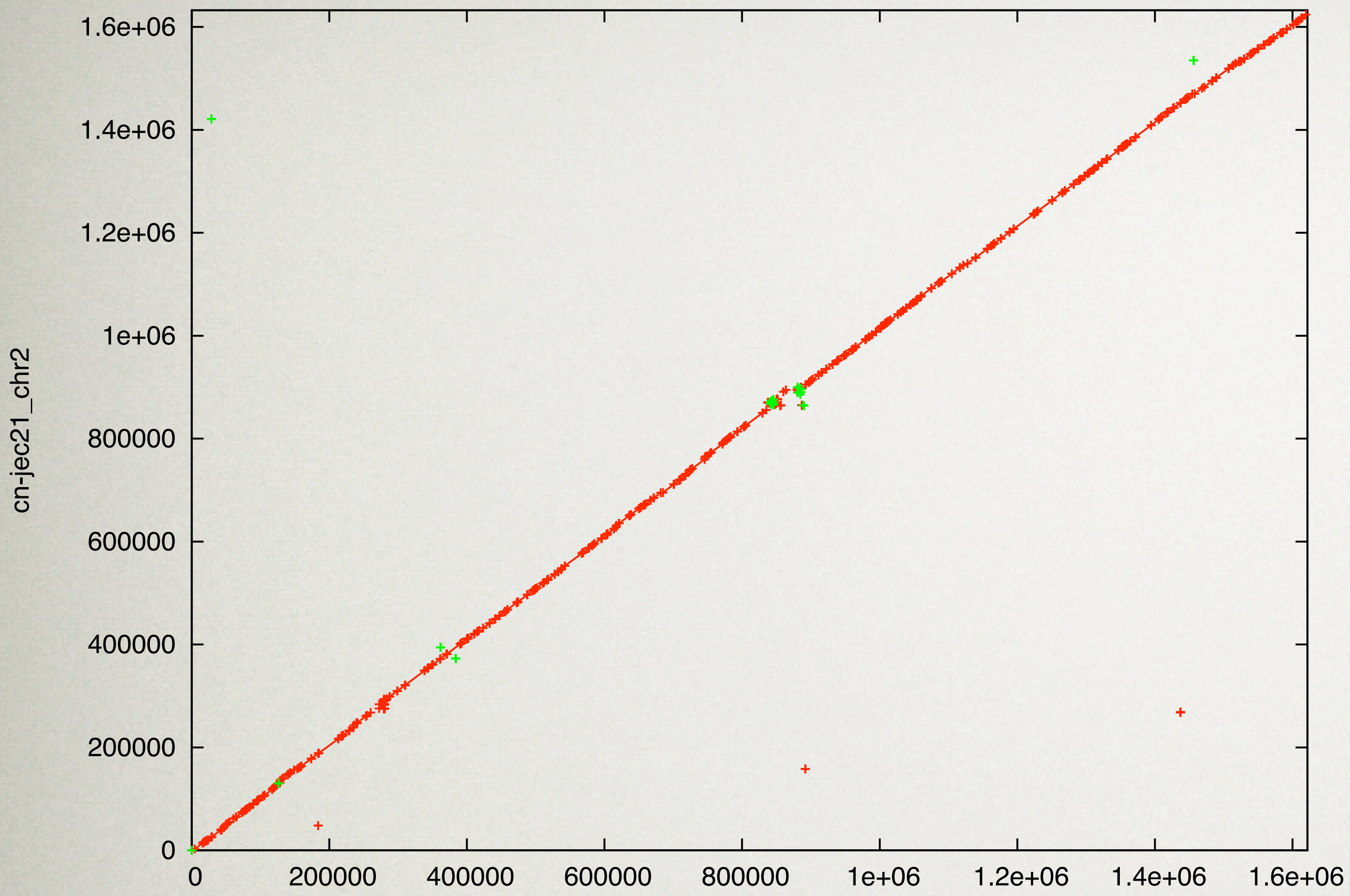
Length = 278,519

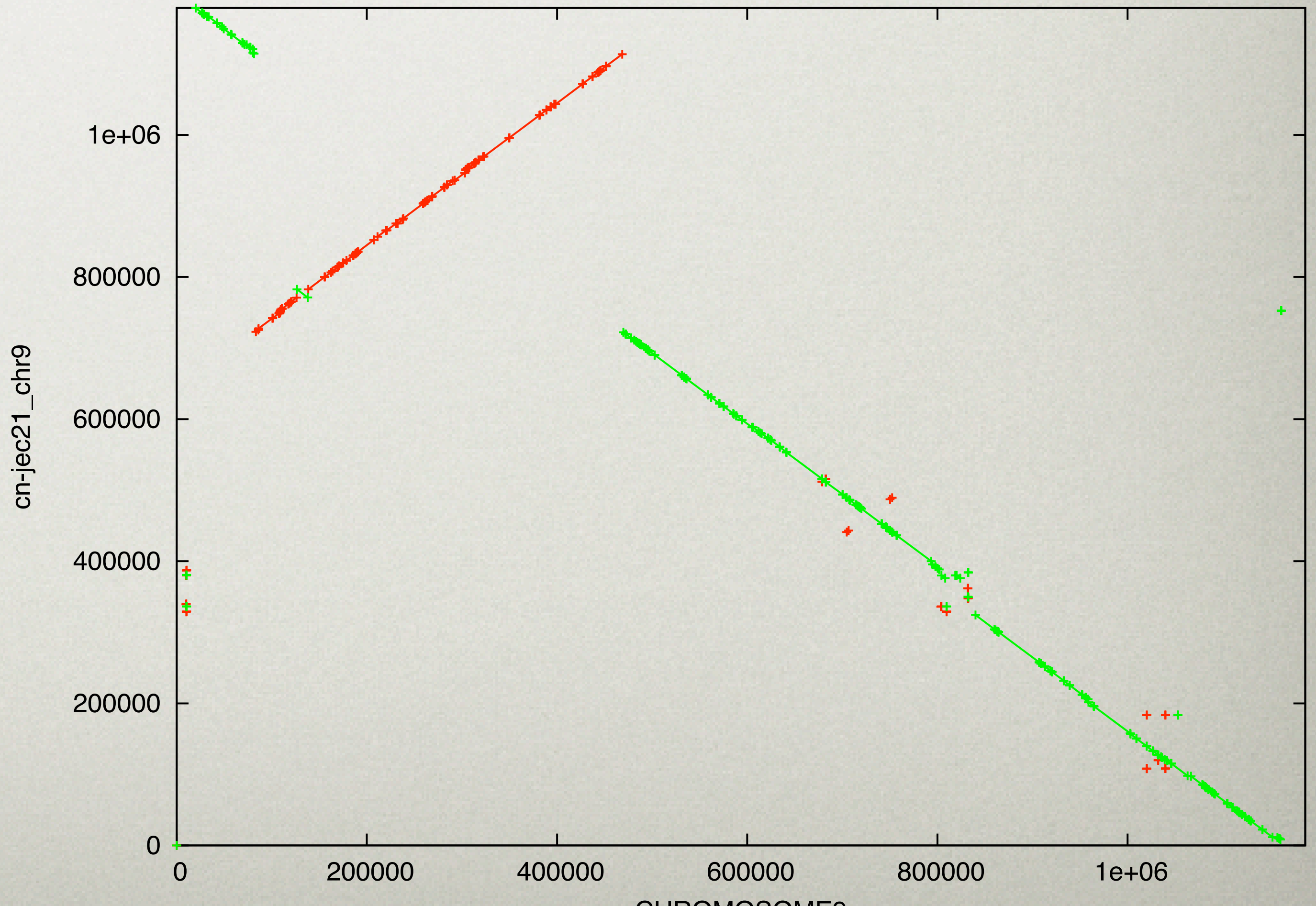
Score = 431.7 bits (1212), Expect = 3.5e-122, P = 3.5e-122
Identities = 222/327 (67%), Positives = 263/327 (80%), Gaps = 1/327 (0%), Frame = +2
Links = (1)

```
Query: 196    FKWVNAEGKFVYIKYHFLADHGQKQFNADEALRYGGEDPDYSKRELWRTIENGKELSWTA 255
              F+ VNAEGKFVY+KYH+LA+HGQKQF   EA+R  GEDPDY+KR+LW  IE G+  +WT
Sbjct: 114953 FRRVNAEGKFVYVKYHYLAEHGQKQFTWPEAVRMSGEDPDYAKRDLWAAIERGETPTWTM 115132


Query: 256    YVQVMKPEDADPEKLGFDPFDVTKVWPKKQFPLQEFGKLTLNKNPENFHRDVEQAAFSPG 315
               VQ+M+PE+ADP KLGFDPFDVTKVWP+ +FP+ E G+L LNKNPEN+HRDVEQ+AFSPG
Sbjct: 115133 KVQIMRPEEADPNKLGFDPFDVTKVWPRSRFPMHEVGRLVLNKNPENYHRDVEQSAFSPG 115312


Query: 316    SMVPGIEDSPDPLLQFRMFFYRDAQYHRIGVNLHQVPVNCPFMASSYSSLNFDGQLRVDA 375
              SMVPGIEDSPD LLQFRMFFYRDAQYHR+GVNLHQ+PVNCPFMA SYSS+NFDG LR DA
Sbjct: 115313 SMVPGIEDSPDALLQFRMFFYRDAQYHRLGVNLHQIPVNCPFMAKSYSSINFDGPLRSDA 115492
```

Comparative Analyses

- ~5553 BRH orthologs between H99 and JEC21
- Some genomic rearrangements, but synteny mostly preserved
- Average Ks 0.22 between A & D across the genome (Mouse/Rat)
    - Ks ~0.35 between A & B or D & B


cn-jec21_chr2
1.6e+06
1.4e+06
1.2e+06
1e+06
800000
600000
400000
200000
0
0
200000
400000
600000
800000
1e+06
1.2e+06
1.4e+06
1.6e+06
CHROMOSOME2


cn-jec21_chr9
1e+06
800000
600000
400000
200000
0
0
200000
400000
600000
800000
1e+06
CHROMOSOME9

Acknowledgments

- Fred Dietrich (Duke)
- Laura Kavanaugh

- Ian Korf (UC Davis)
- Aaron Mackey (U Penn)
- Duke CSEM Cluster
- Broad, TIGR, BC GSC