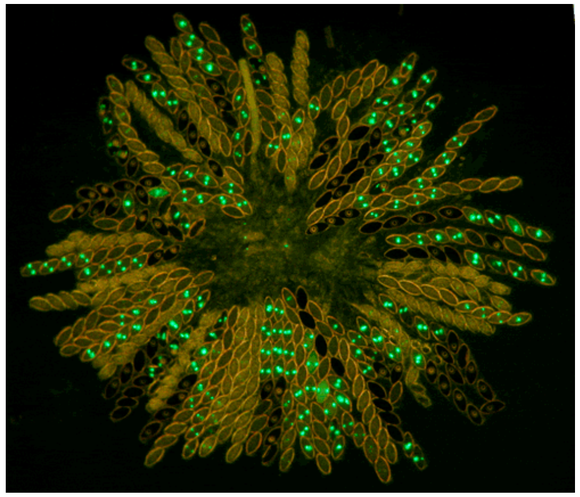
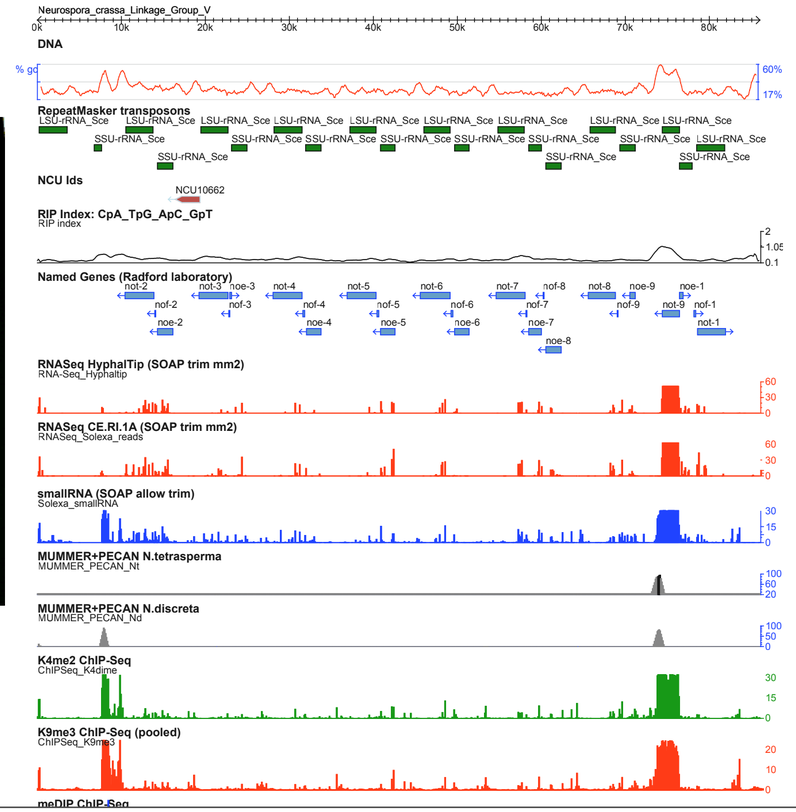
Profiling evolution in the fungus *Neurospora crassa* with transcriptional and comparative genomics

Jason Stajich
Plant and Microbial Biology
University of California, Berkeley

Plant Pathology and Microbiology
University of California, Riverside





Overview

- Introducing the system: *Neurospora*
- Sequence data collection
- Comparative genomics
- Describing the transcriptome
  - Transcriptional Profiling with next generation sequencing
  - SmallRNA profiling

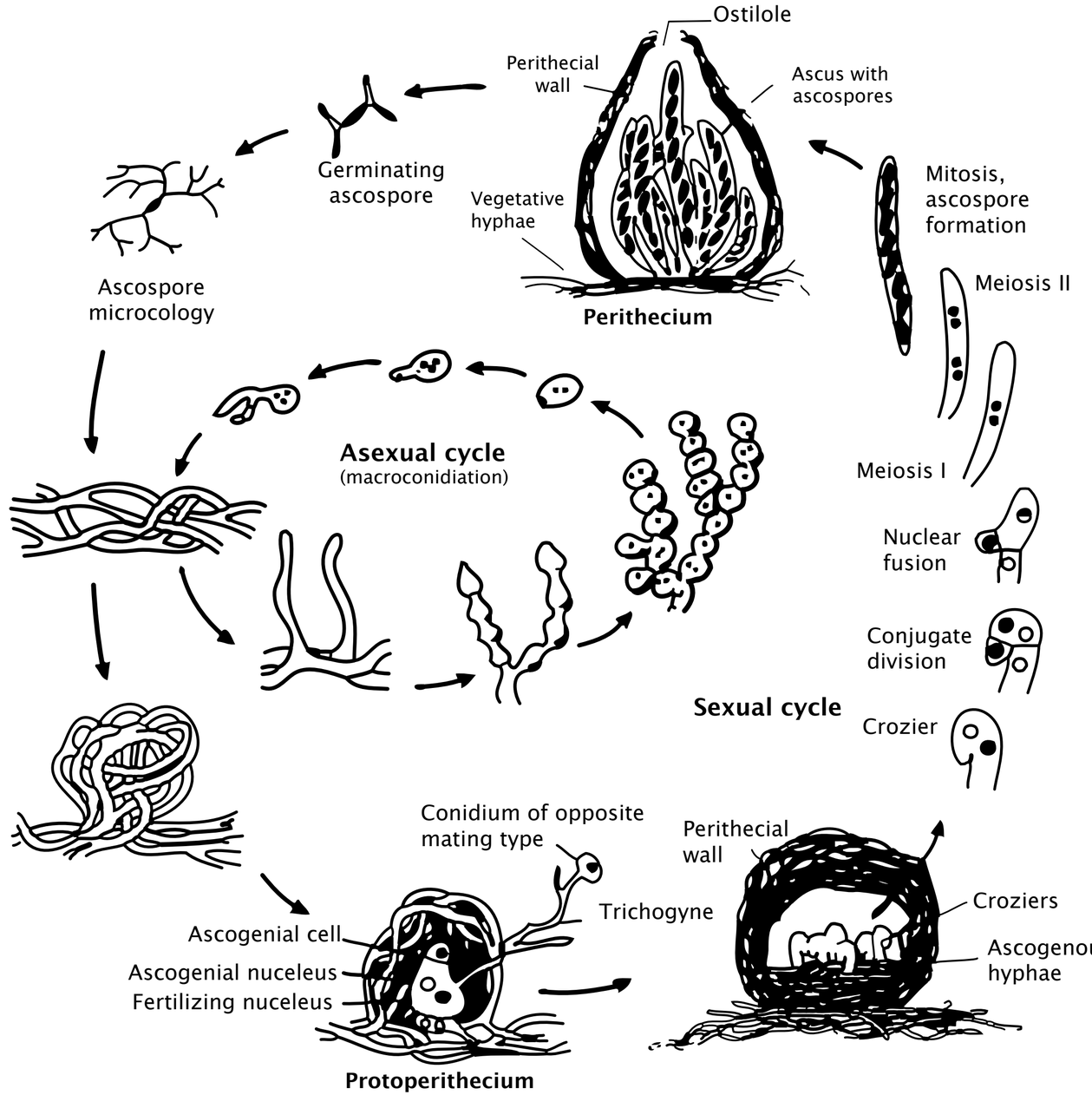
Neurospora

- A filamentous fungus that is an important model (Reference) organism for genetic and developmental studies

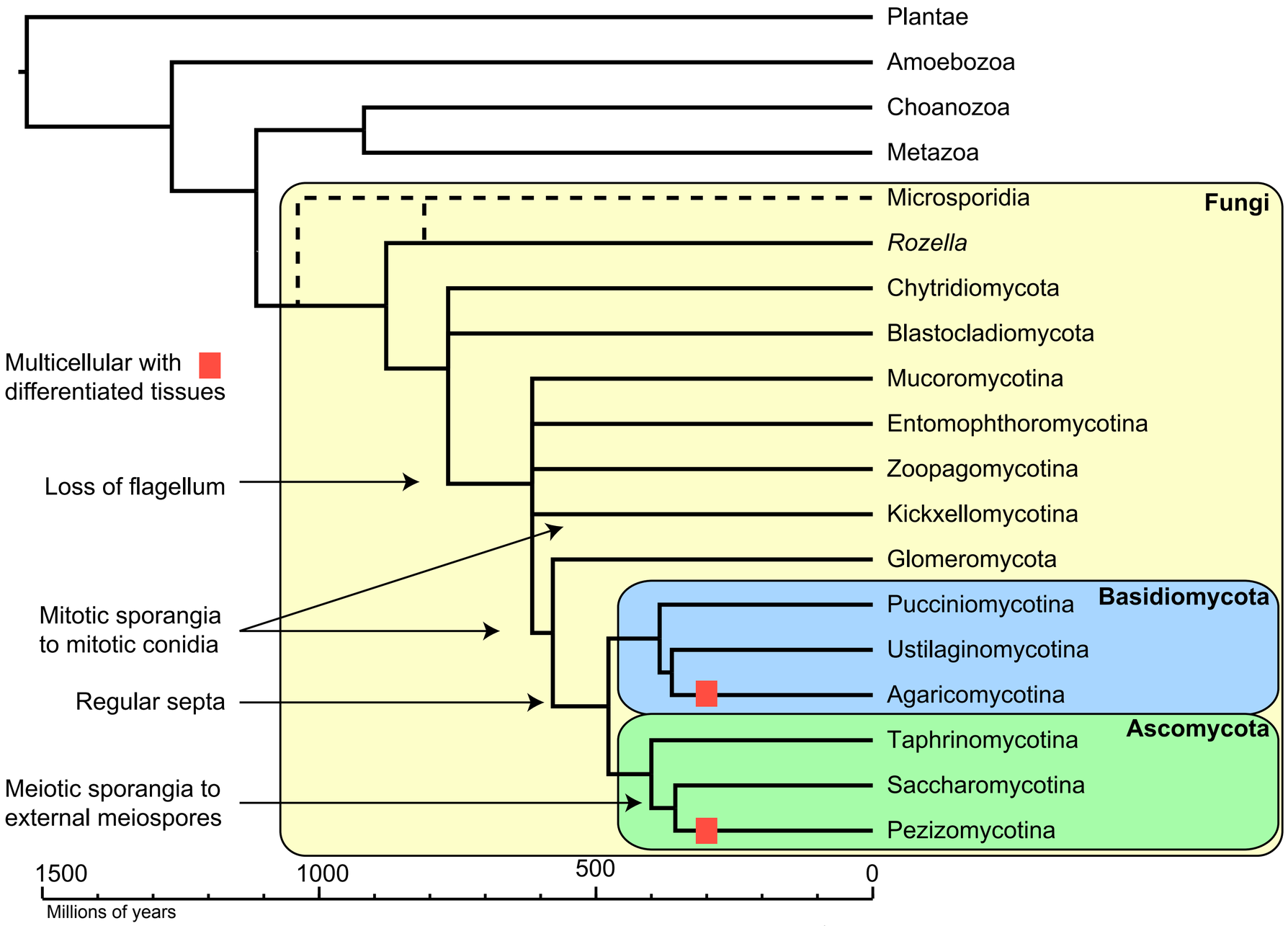
- Has a well-sequenced genome with ~10,000 genes

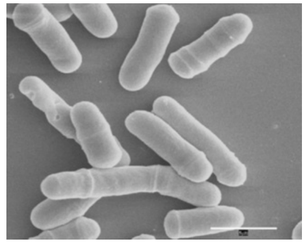
- Systematic knockout project with more than 60% of the genes knocked out for both mating types

- Evolutionary study system due to extensive sampling of related species



From The Biology of Neurospora R. Davis

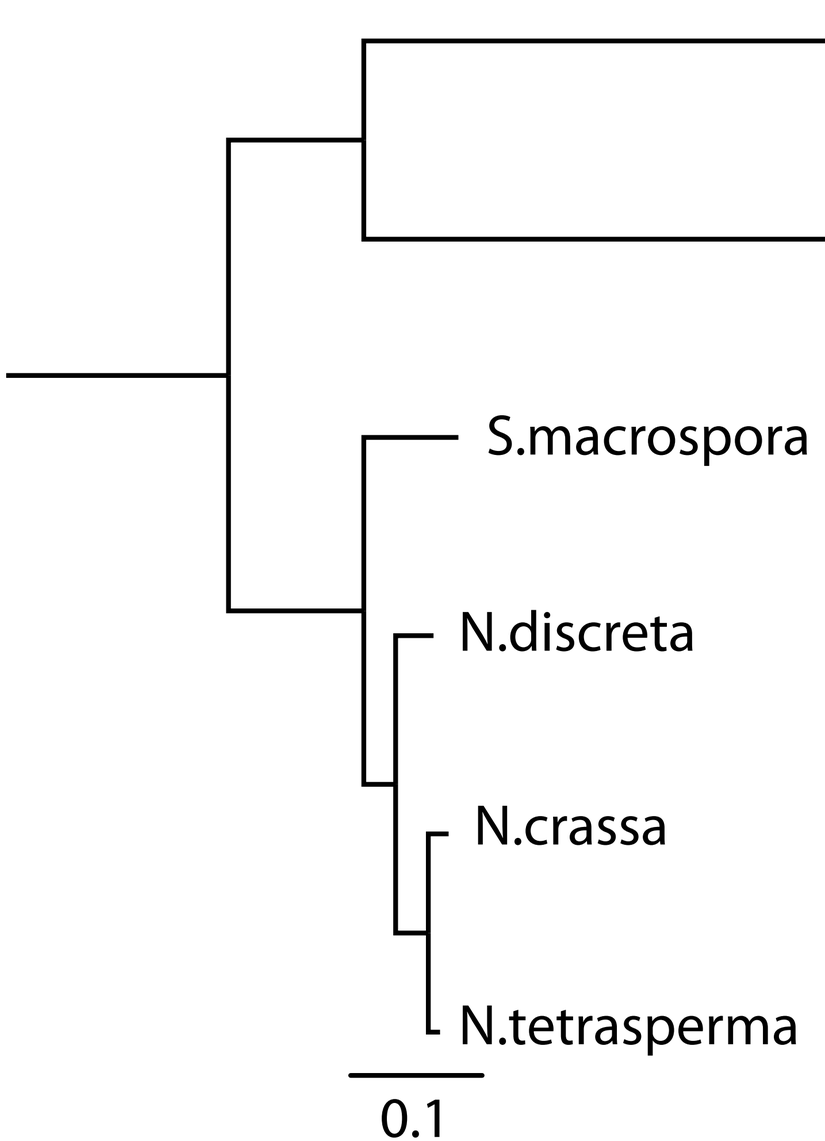




Fungal Phylogeny of Major branches

Stajich et al, *Current Biol in press*
based on James et al, *Nature* 2006.

Sordariales genome sequencing projects



CNRS/Paris-Sud

Broad

Solexa + 454 (Kueck, Freitag)

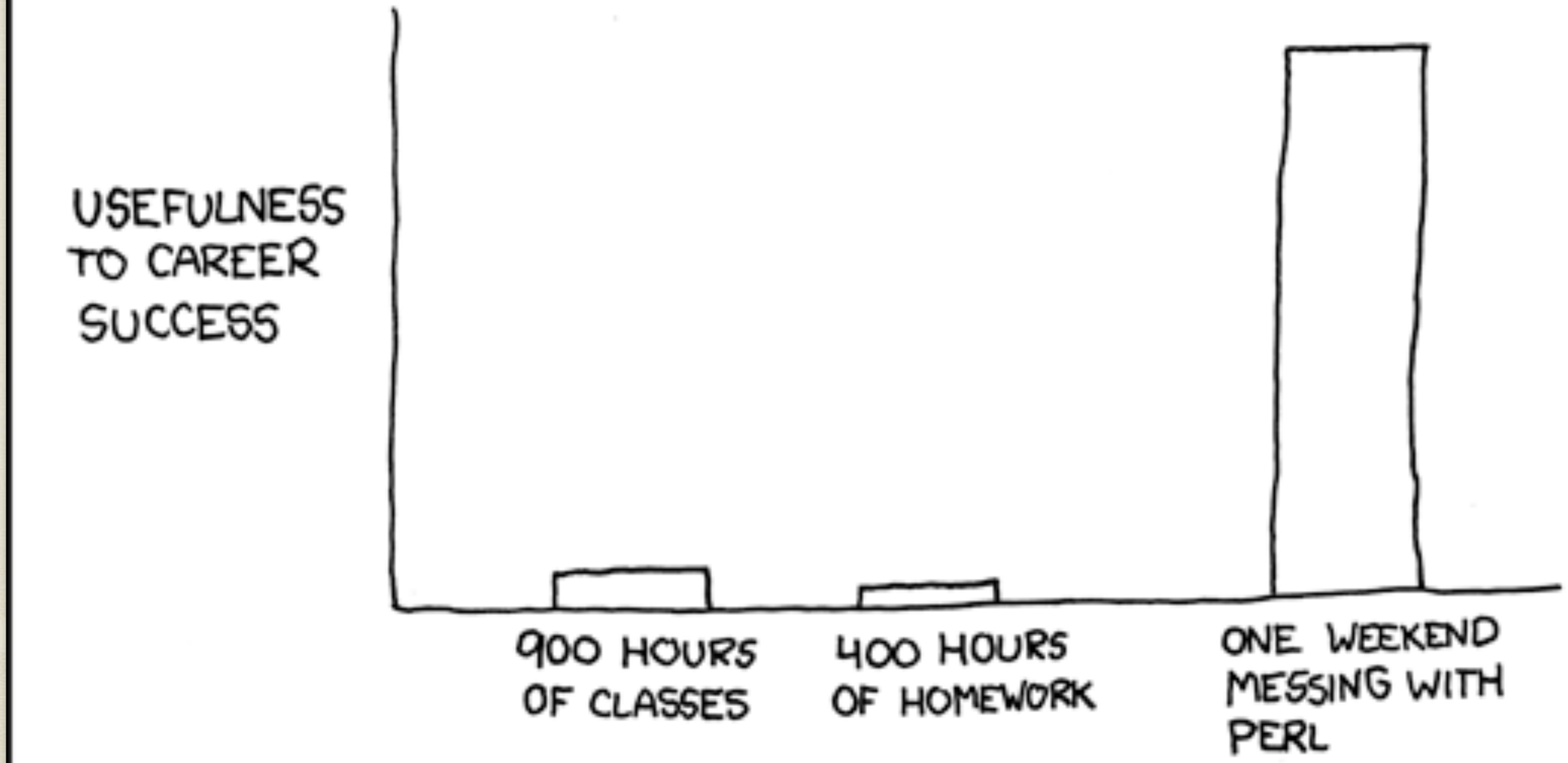
JGI

Broad

100+ Nc strains for Association project (Brem, Glass, Taylor)

JGI

College, Gradschool ~~11TH~~-GRADE ACTIVITIES:



xkcd.org

Some fungal genome research questions

- How did the morphological complexity of the fungi evolve?
  How do species form in fungi?

- How does genome structure change over time? Are these changes neutral or selective?

  - How many rearrangements are there between species? How do these contribute to formation of biological species boundaries (failure to hybridize).

  - What are the (relatively) fast and what are slow evolving parts of the genome

- What is the complete gene set of a filamentous fungus?

  - First defining the set of genes- protein coding and non protein-coding genes

  - Comparative genomics to discover which genes were gained or lost. Does this relate to the evolution of yeast or filamentous only-forms?

- How are genes regulated to form different tissues or forms in fungi?

  - Comparing gene expression among different stages?
    What are the master regulators?

Some fungal genome research questions

- How did the morphological complexity of the fungi evolve?
  How do species form in fungi?

- How does genome structure change over time? Are these changes neutral or selective?

  - How many rearrangements are there between species? How do these contribute to formation of biological species boundaries (failure to hybridize).

  - What are the (relatively) fast and what are slow evolving parts of the genome

- What is the complete gene set of a filamentous fungus?

  - First defining the set of genes- protein coding and non protein-coding genes

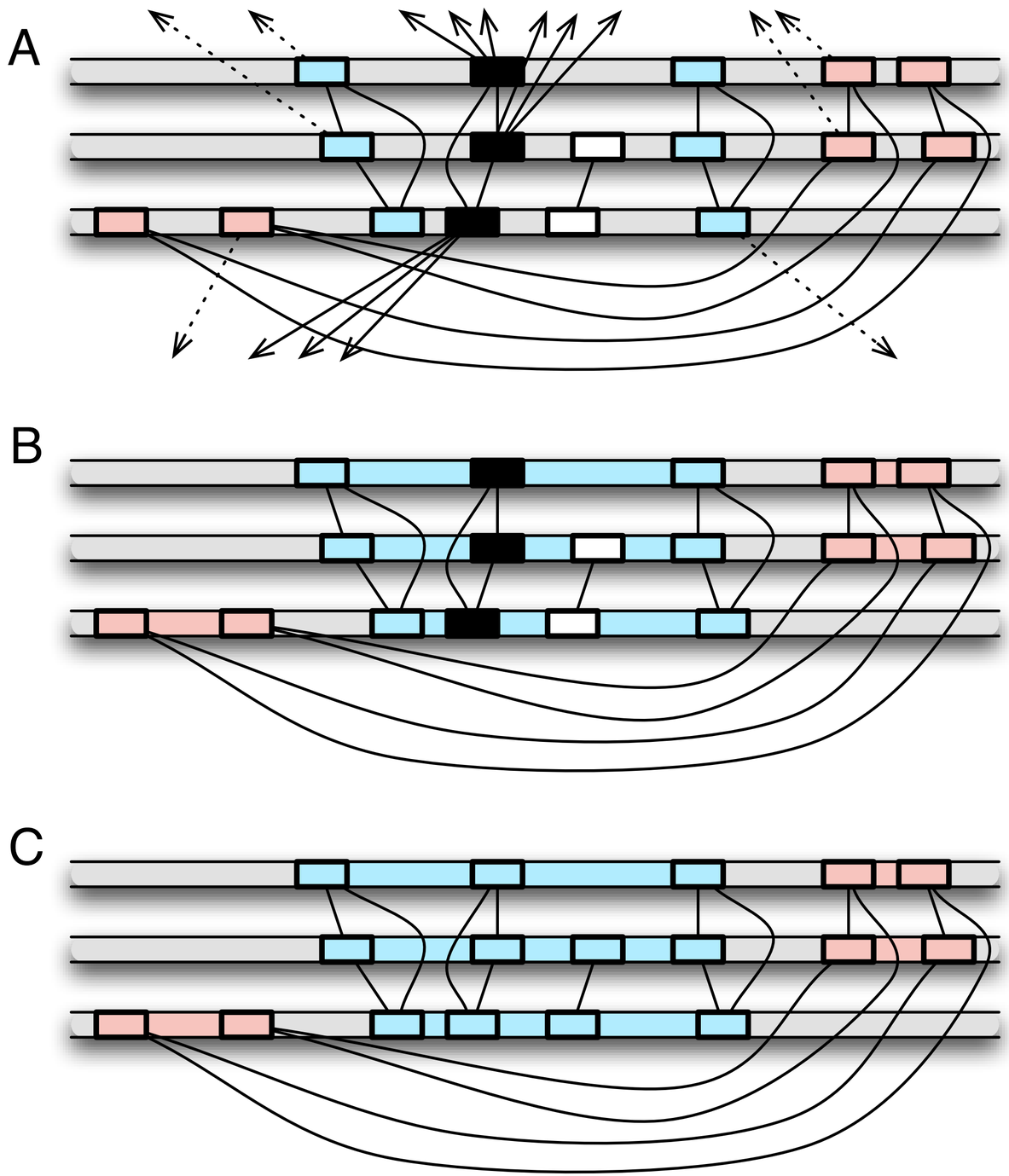
  - Comparative genomics to discover which genes were gained or lost. Does this relate to the evolution of yeast or filamentous only-forms?

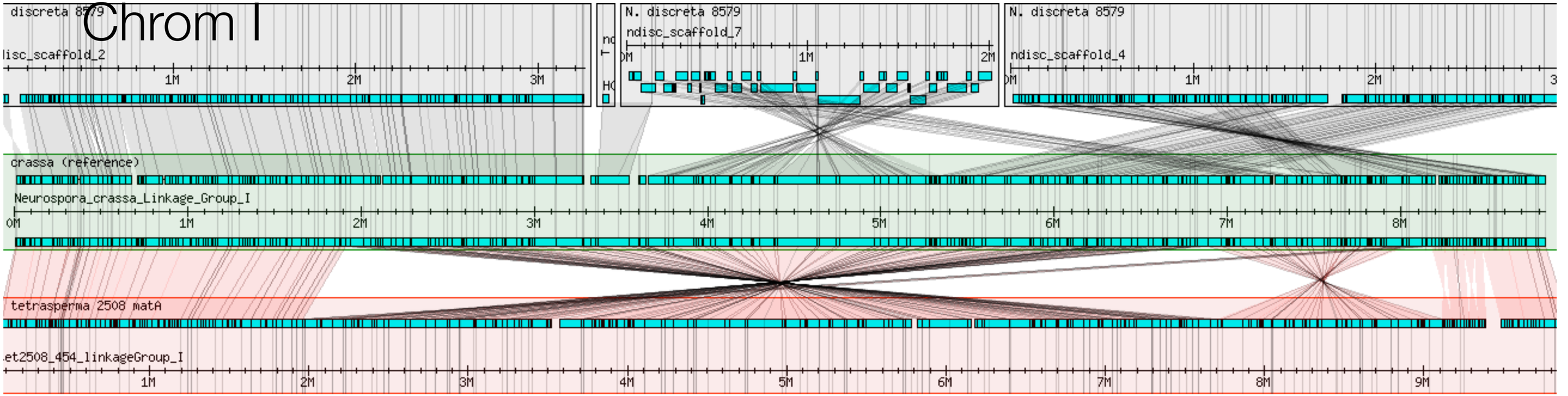
- How are genes regulated to form different tissues or forms in fungi?

  - Comparing gene expression among different stages?
    What are the master regulators?

Synteny with Mercator

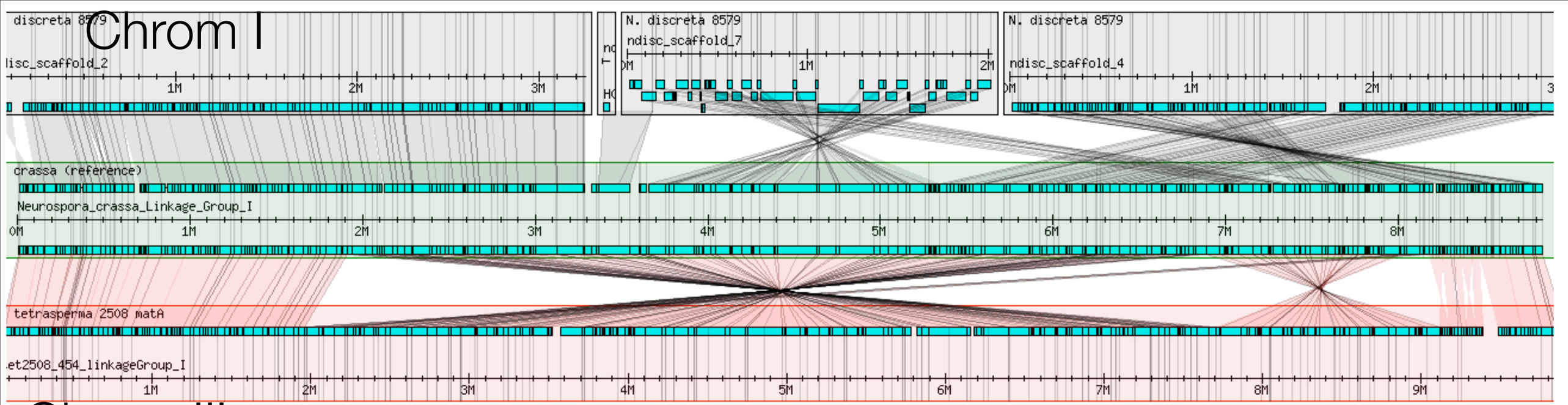
- Identify anchors - similar regions in both genomes.
    - Exons or just highly similar nucleotide regions between species.
- Link anchors into logical groups
- Identify boundaries of the block via alignments
- Align blocks with multiple alignment tool (i.e. Clustalw, MAVID, PECAN)



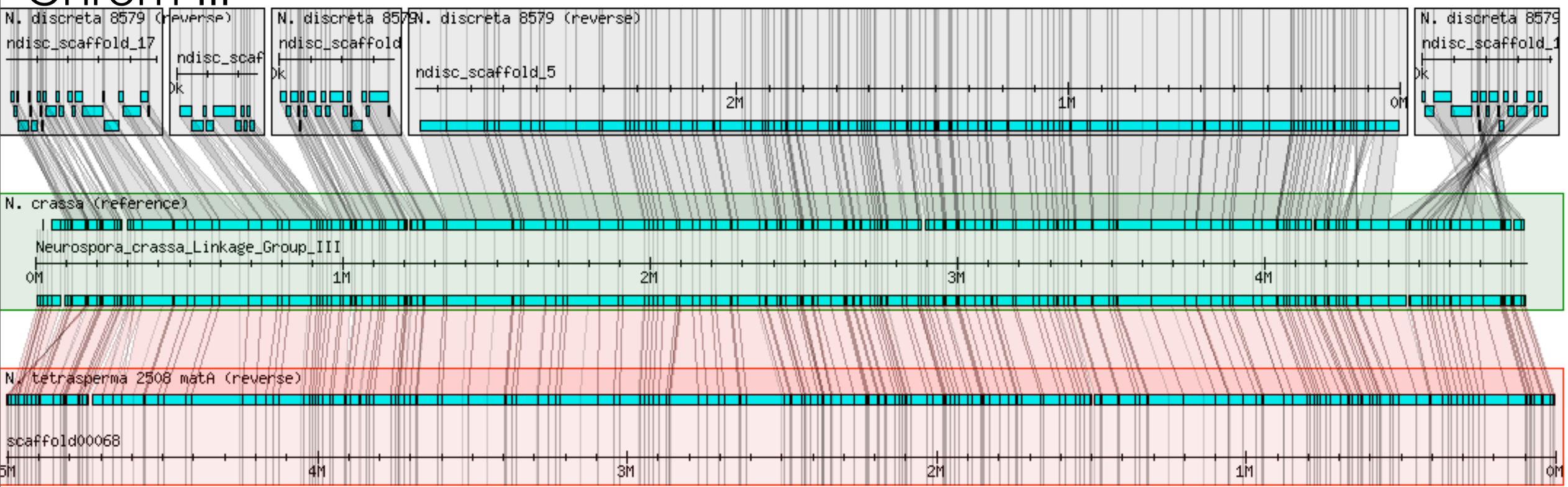


Neurospora synteny

Chrom I



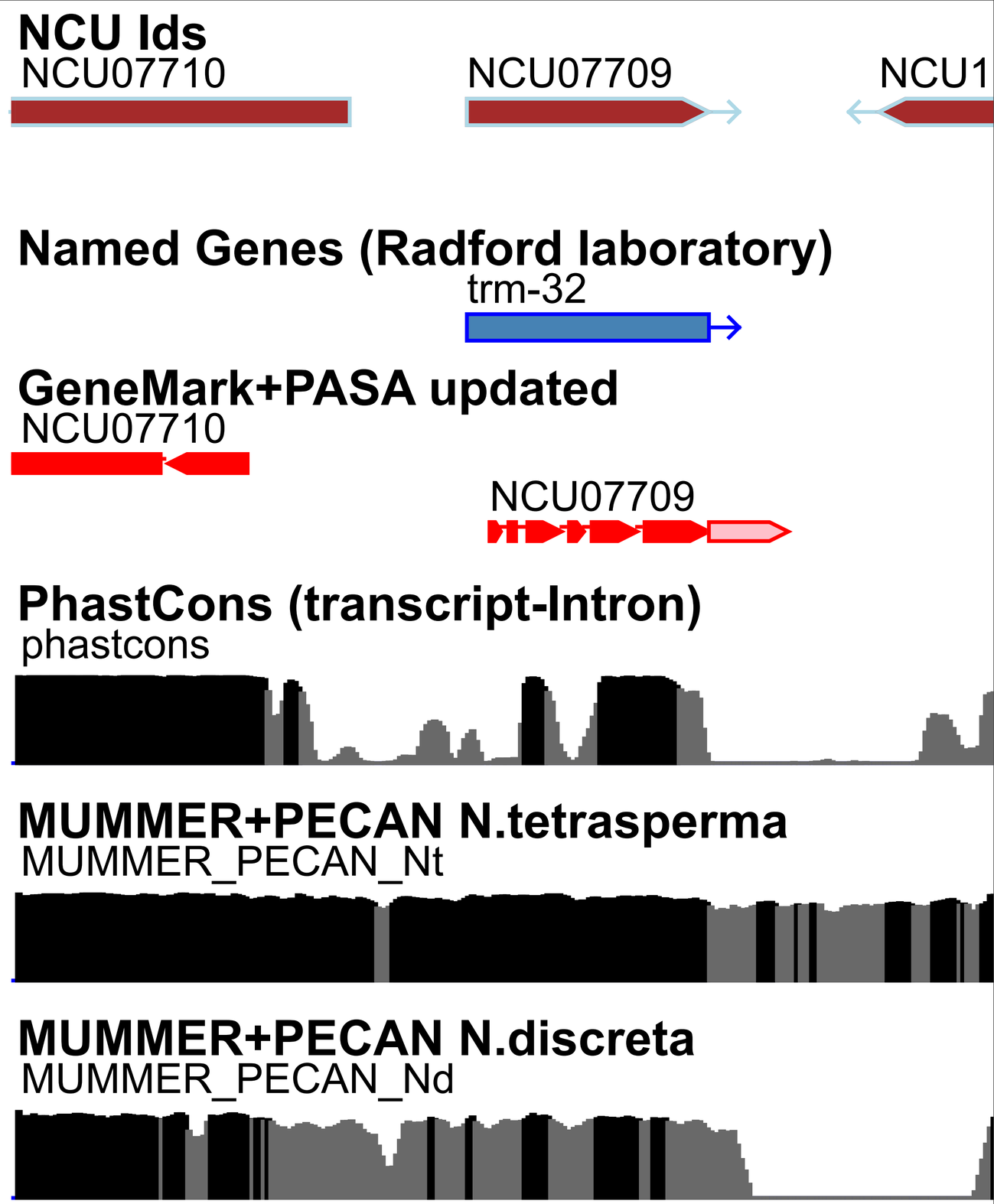
Chrom III

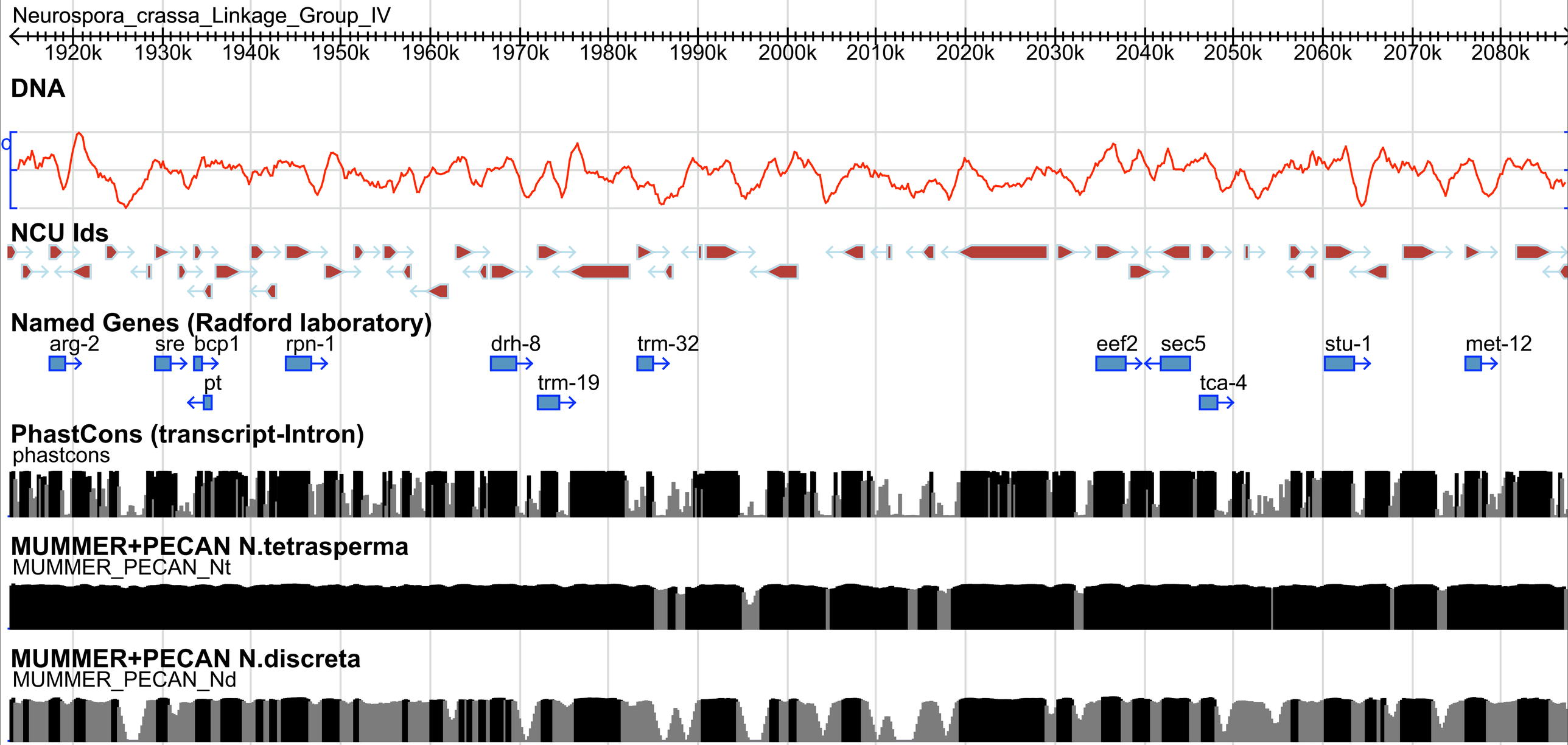


Neurospora synteny

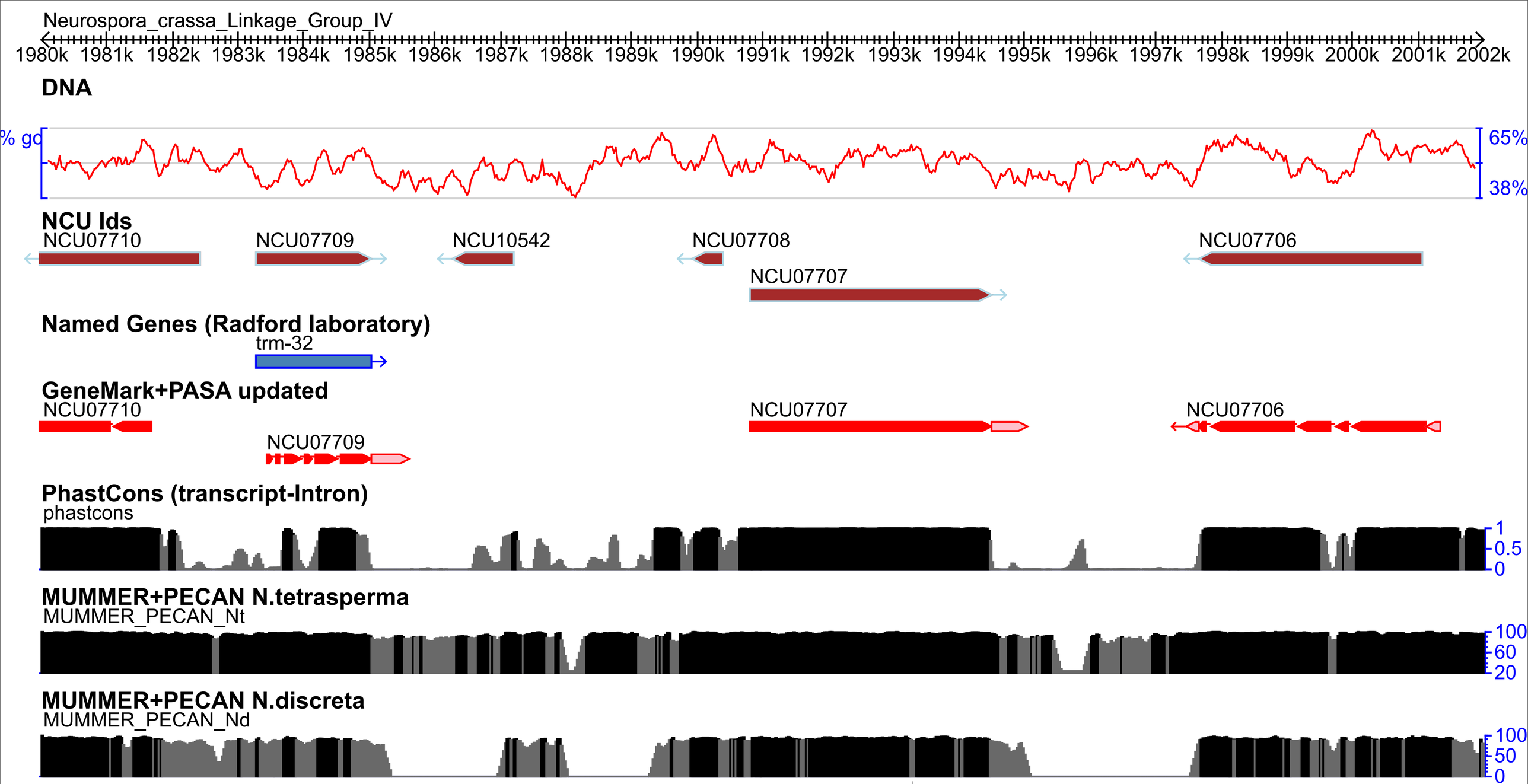
Nucleotide conservation

- Simple percent identity calculation
- PhastCons two-state HMM for constraint at basepair level. Trained on Exons vs Introns
- Run whole genome 3-way alignments to identify constrained regions
- Identify novel conserved regions and patterns of evolution on different feature types





PhastCons and %id conservation



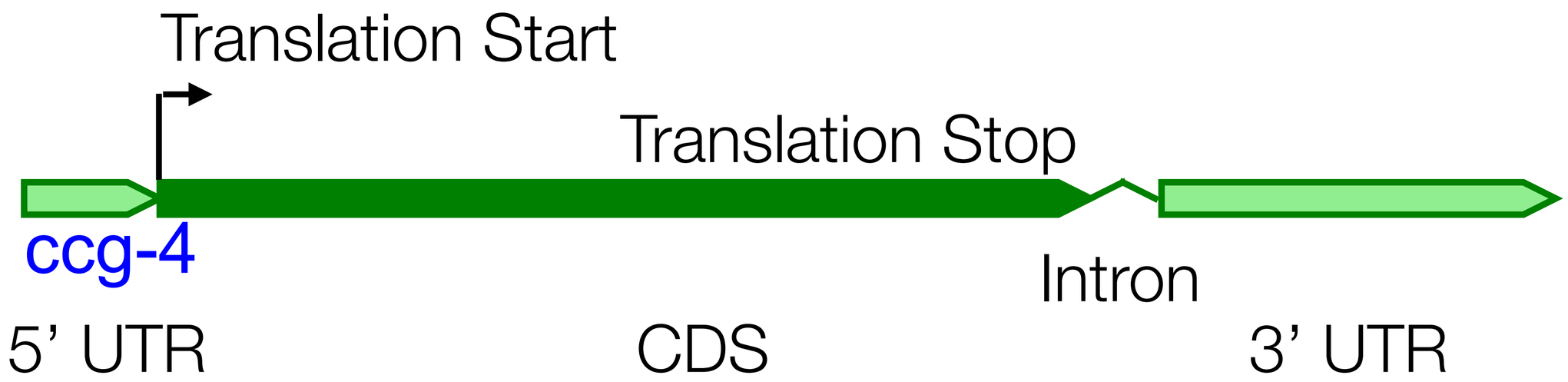
PhastCons and %id conservation

!Next gen sequencing will/has change(d) Biology!

- Believe the hype?
    - LOTS more data is and will be available. How to integrate it?
    - RNA-Seq - sequencing RNA, randomly sheared or biased towards 3'-end mRNA (poly-A purified)
      ChIP-Seq - sequencing DNA that is bound by proteins that are purified by antibody pulldown
      smallRNA-Seq - sequencing smallRNA fraction of the transcriptome
- You will hear about a lot of different short-read mappers. Advantages and disadvantages to many. For these data I have used SOAP and BowTie+TopHat
    - SOAP for small RNA reads since length of reads can vary
      BowTie+TopHat for for mapping of RNASeq reads

Improving the genome annotation

- Expressed Sequence Tag sequencing with longer read technology (454) and Solexa RNA-Seq
- Comparative genomics to find conserved genomic regions
- Identify new exons, splice-sites, UnTranslated (UTR) regions
- Identify new genes



“Traditional” mRNA sequencing - RNA-Seq

- 977k ESTs from *N. crassa* and related species
  251k Sanger EST sequences for *N. crassa*
  454 sequencing:
  450k from *N. discreta*
  273k from *N. tetrasperma*

- ESTs aligned with splicing to the genome and new gene calls made using Sim4 and Gmap (PASA; Haas et al. 2003)

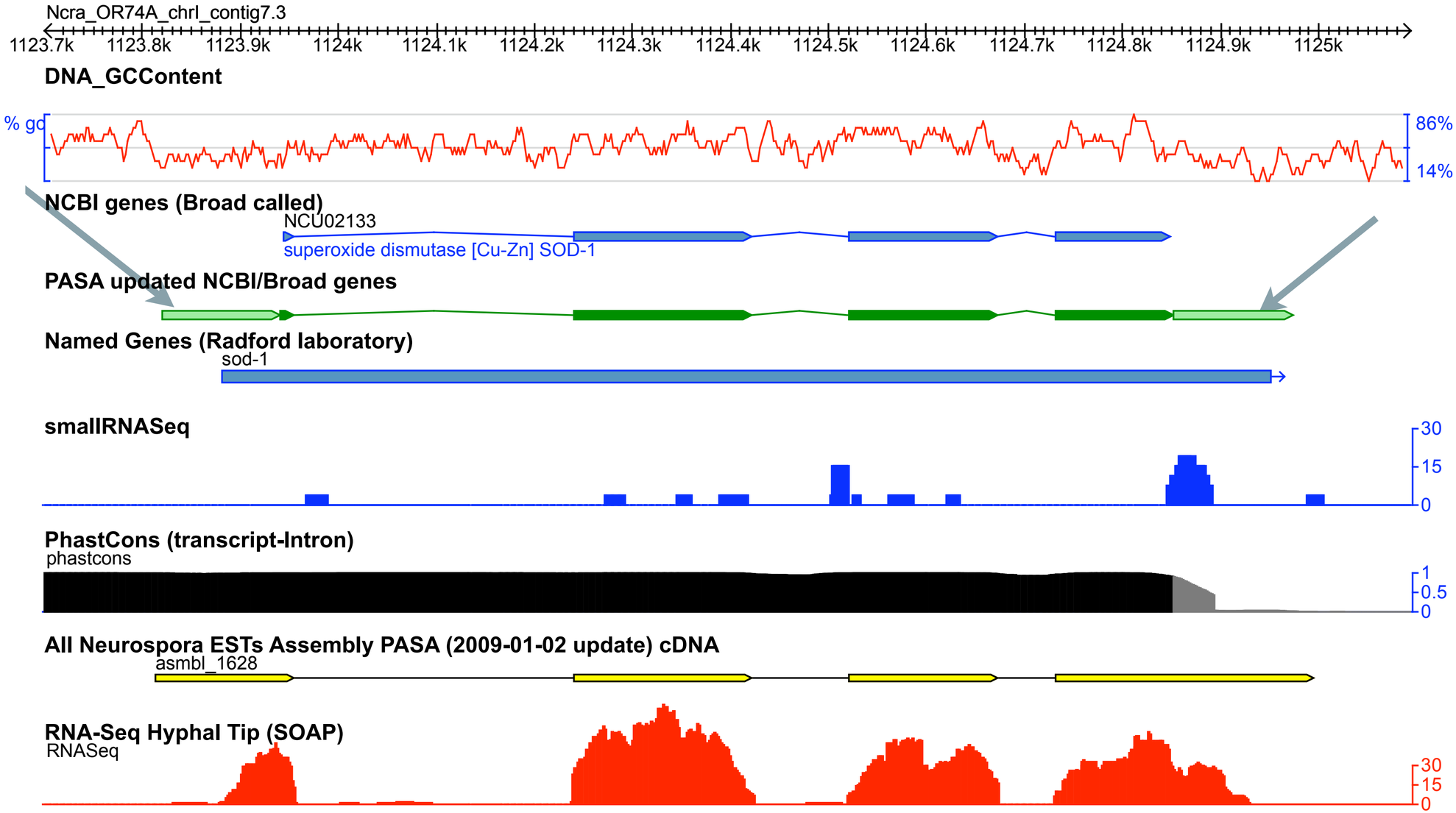
- Coverage of 7,496 genes with successful incorporation of ESTs (80%)
  1,941 genes with conflict or no-EST support (20%)

- No UTRs in published *Neurospora* annotation

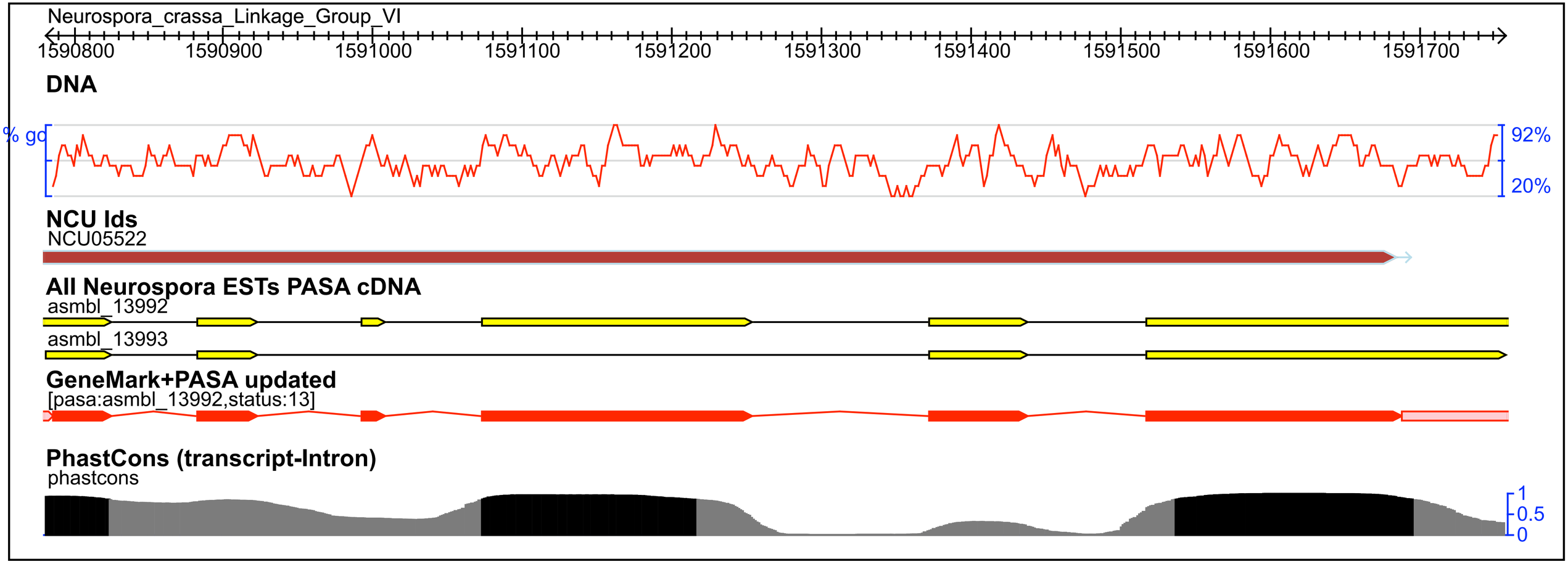
- Update Gene annotations:
  5,311 genes with 5’ UTR 6,275 genes 3’ UTRs
  378 alternative splicing events, but only 7 isoforms with alternative exon content

Updated annotation using ESTs

5'UTR
5,311
genes
55%



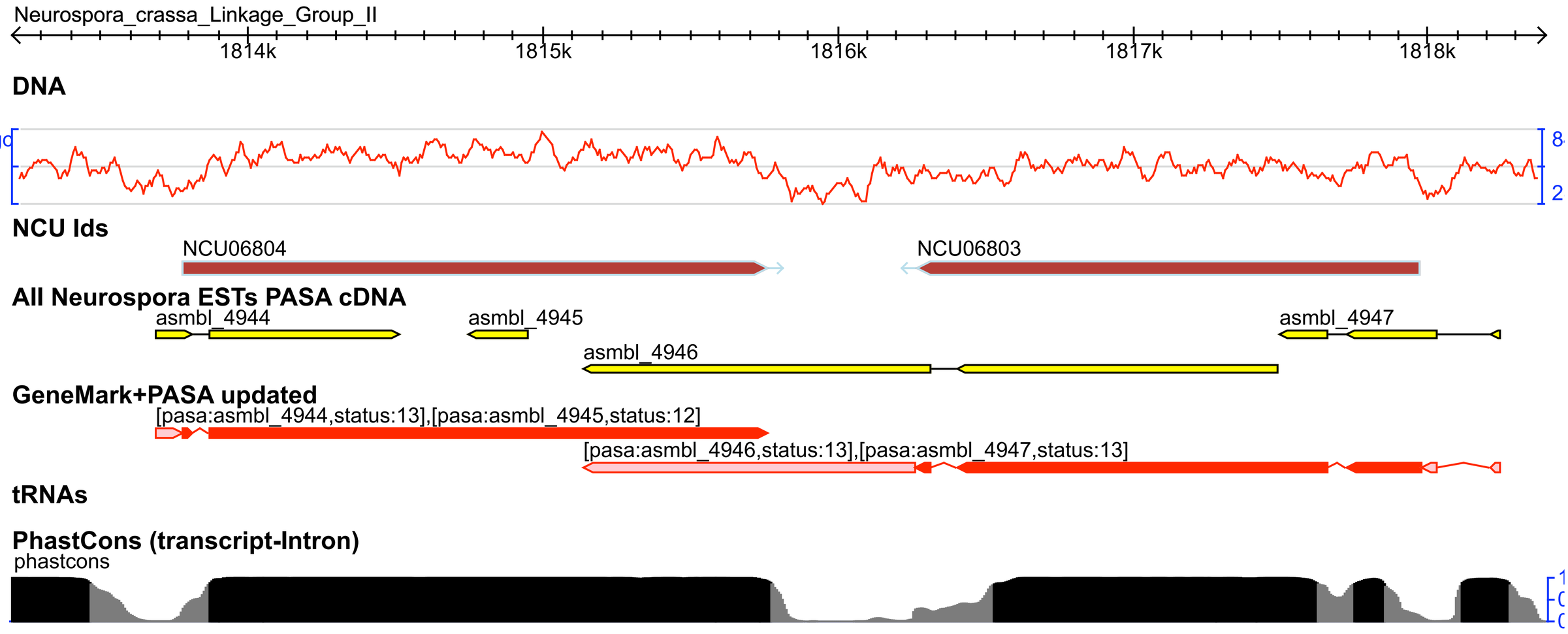
3'UTR
6,275
genes
65%



Almost 80 candidates loci with exon skipping or alternative inclusion from the ESTs (PASA)

Alternative splicing

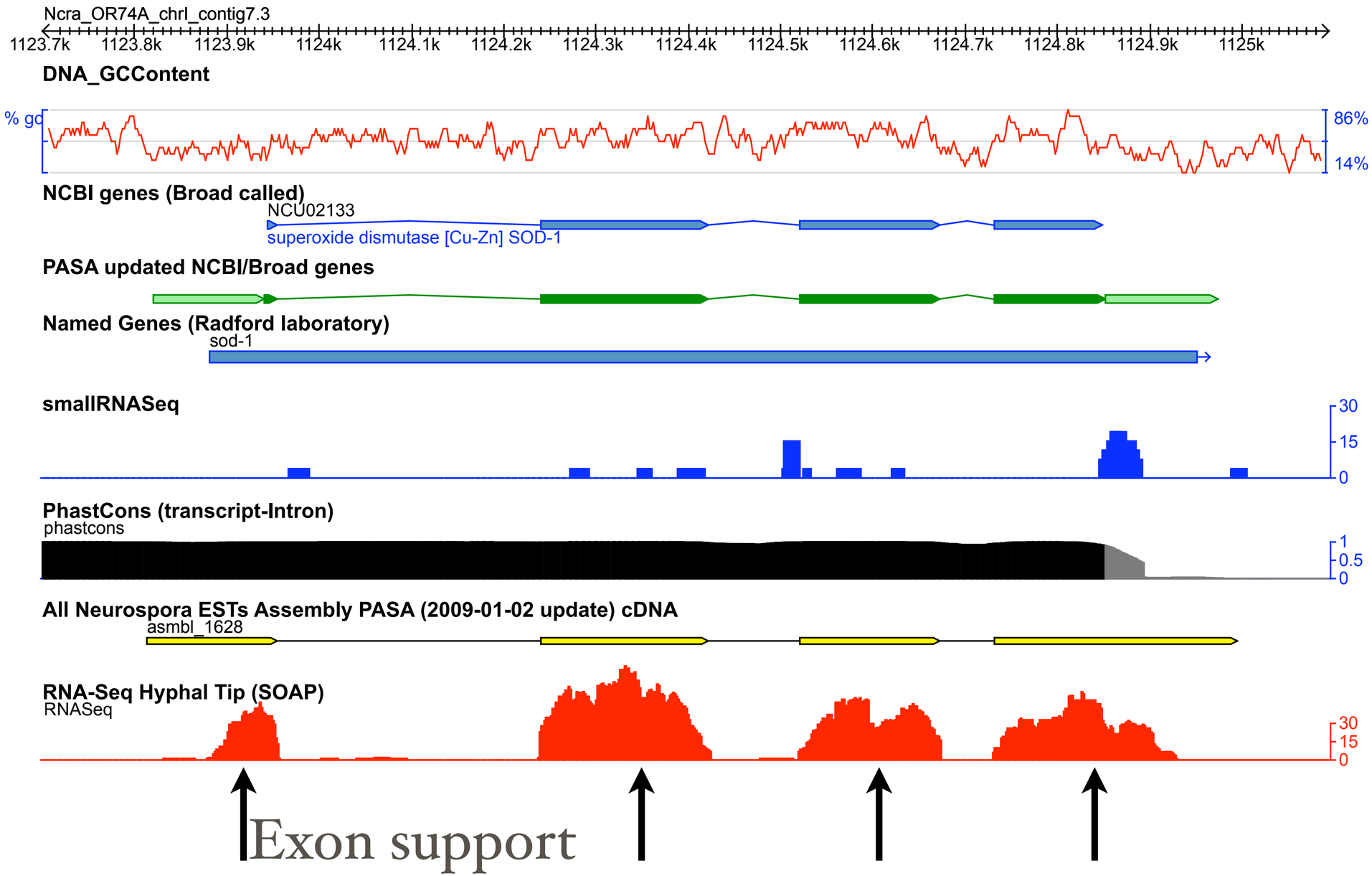
Overlapping genes



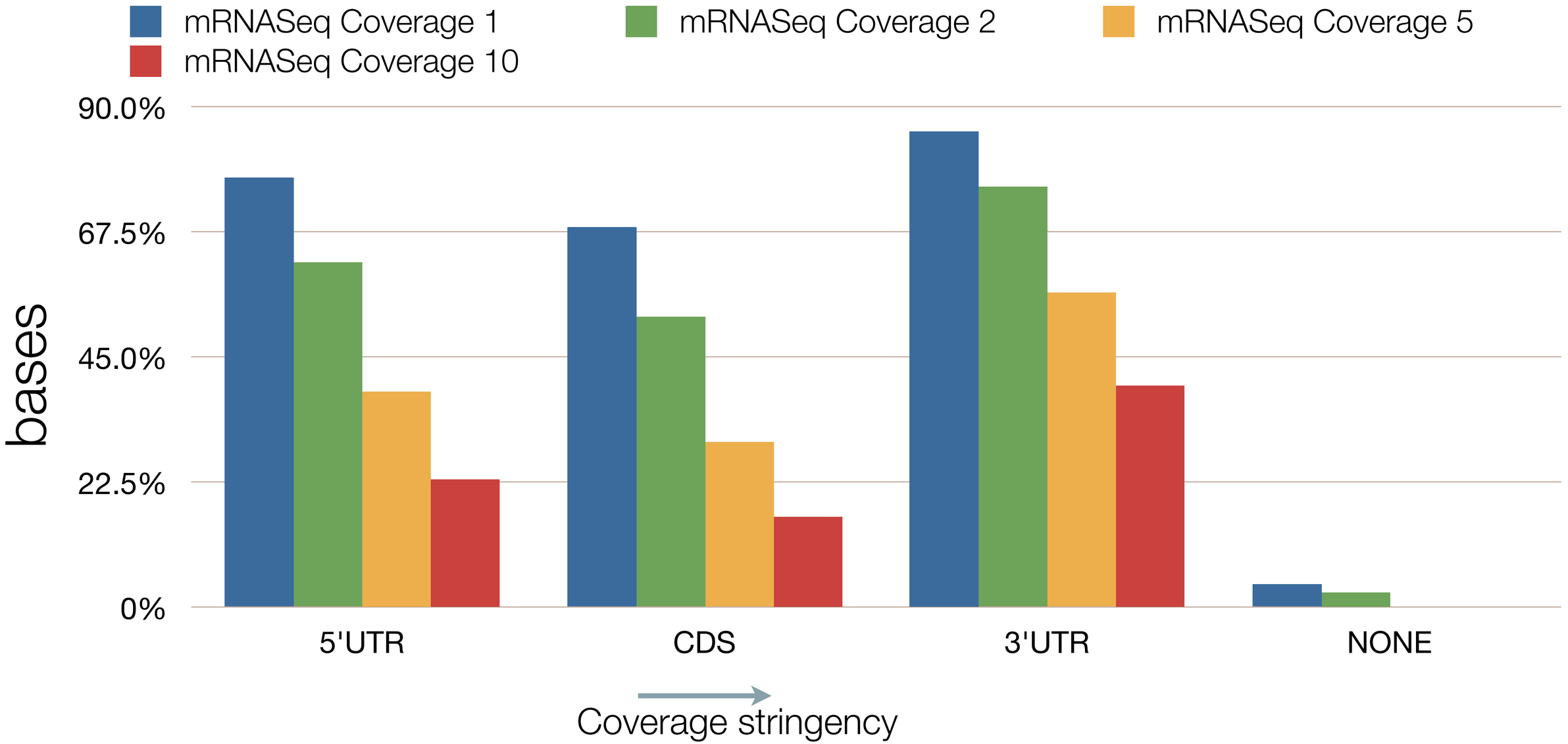
~200 convergently transcribed genes overlap, mostly in 3' UTR

RNA-Seq in *Neurospora crassa*

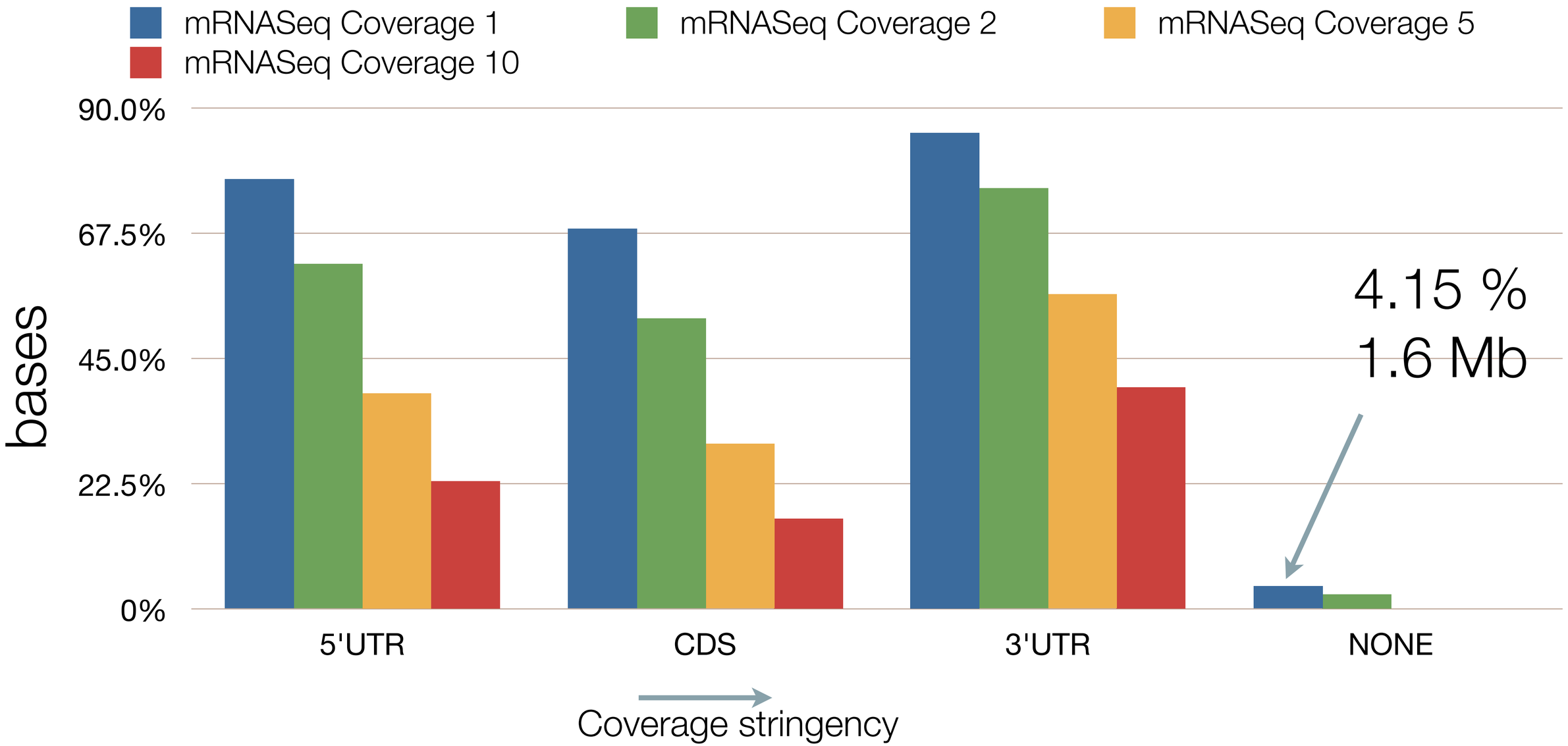
RNA-Seq mapped to the genome



mRNASeq coverage of gene regions



mRNASeq coverage of gene regions

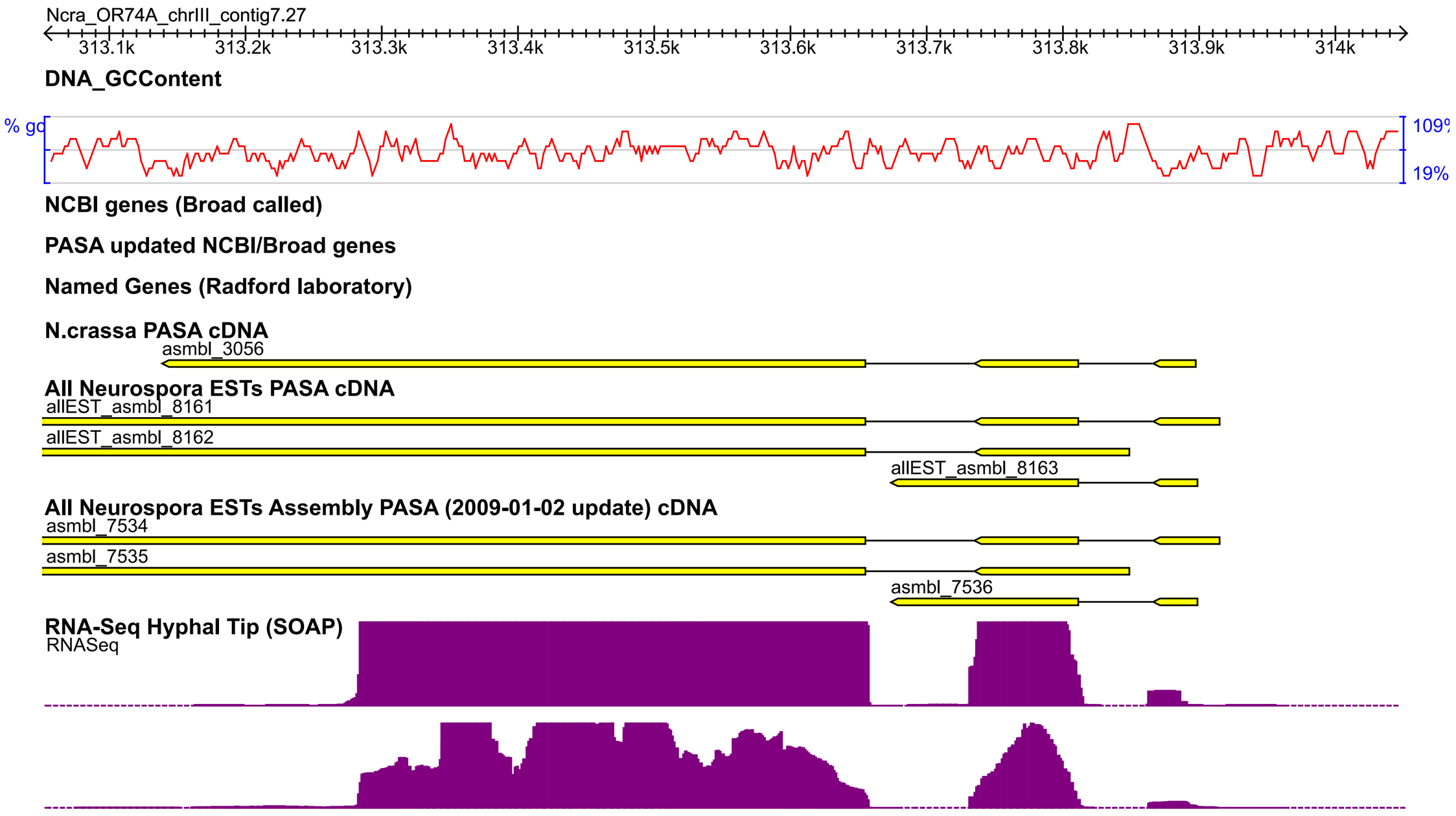


Finding new genes

- ~600 new splice-site regions in the genome
    - Identify new UTR sequences
- ~170 new gene regions previously not overlapping genes

- Can use RNA-Seq to flag candidate regions with transcription
    - By requiring splicing limits predicted regions processed transcripts rather than ectopic or random transcription
    - Further refine gene models to include additional genes
    - Potential identification of noncoding RNAs through these regions

New Genes


Ncra_OR74A_chrIII_contig7.27
313.1k
313.2k
313.3k
313.4k
313.5k
313.6k
313.7k
313.8k
313.9k
314k
DNA_GCContent
% gc
109%
19%
NCBI genes (Broad called)
PASA updated NCBI/Broad genes
Named Genes (Radford laboratory)
N.crassa PASA cDNA
asmbl_3056
All Neurospora ESTs PASA cDNA
allEST_asmbl_8161
allEST_asmbl_8162
allEST_asmbl_8163
All Neurospora ESTs Assembly PASA (2009-01-02 update) cDNA
asmbl_7534
asmbl_7535
asmbl_7536
RNA-Seq Hyphal Tip (SOAP)
RNASeq

RNA-Seq vs “traditional” EST seq

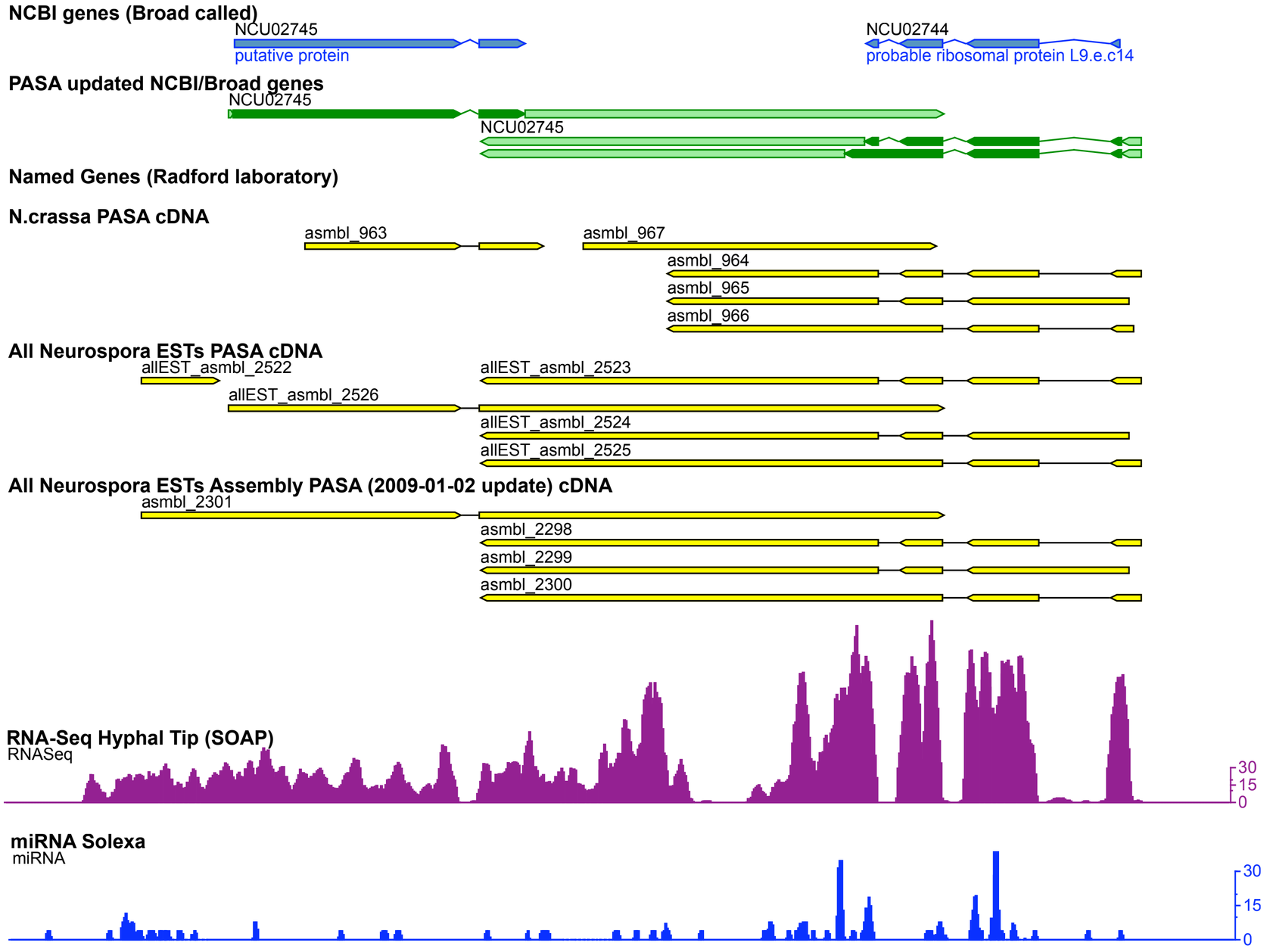
- RNA-Seq - 90% of genes have RNA-Seq transcripts at any part of gene
  (1 lane of 1 flowcell ~$800-1000)

- ESTs - 80% of genes have transcripts
  900K ESTS including ~600k 454 ESTs at cost of ~$25-30k

- Still validating the predicted splice-sites from both sets.
  Little alternative splicing from the 454 data, but limited conditions. Methodology for RNA-Seq/Illumina data of 35-40bp still being worked out or superceeded by longer read libraries.

- Expression varies among developmental conditions so pooling or multiple RNA-Seq libraries will be essential to fully describe transcriptome.

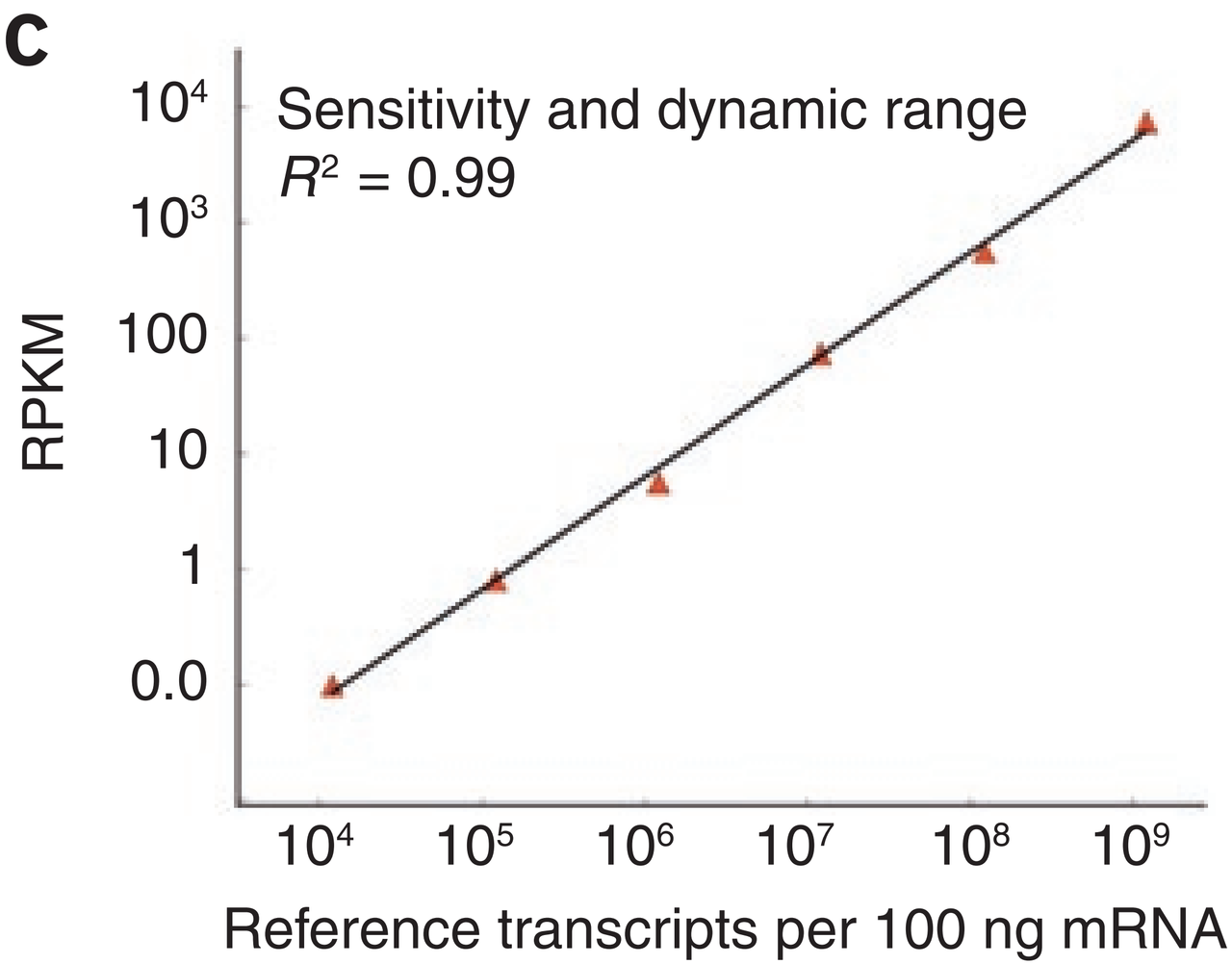
Overlapping genes & small RNAs



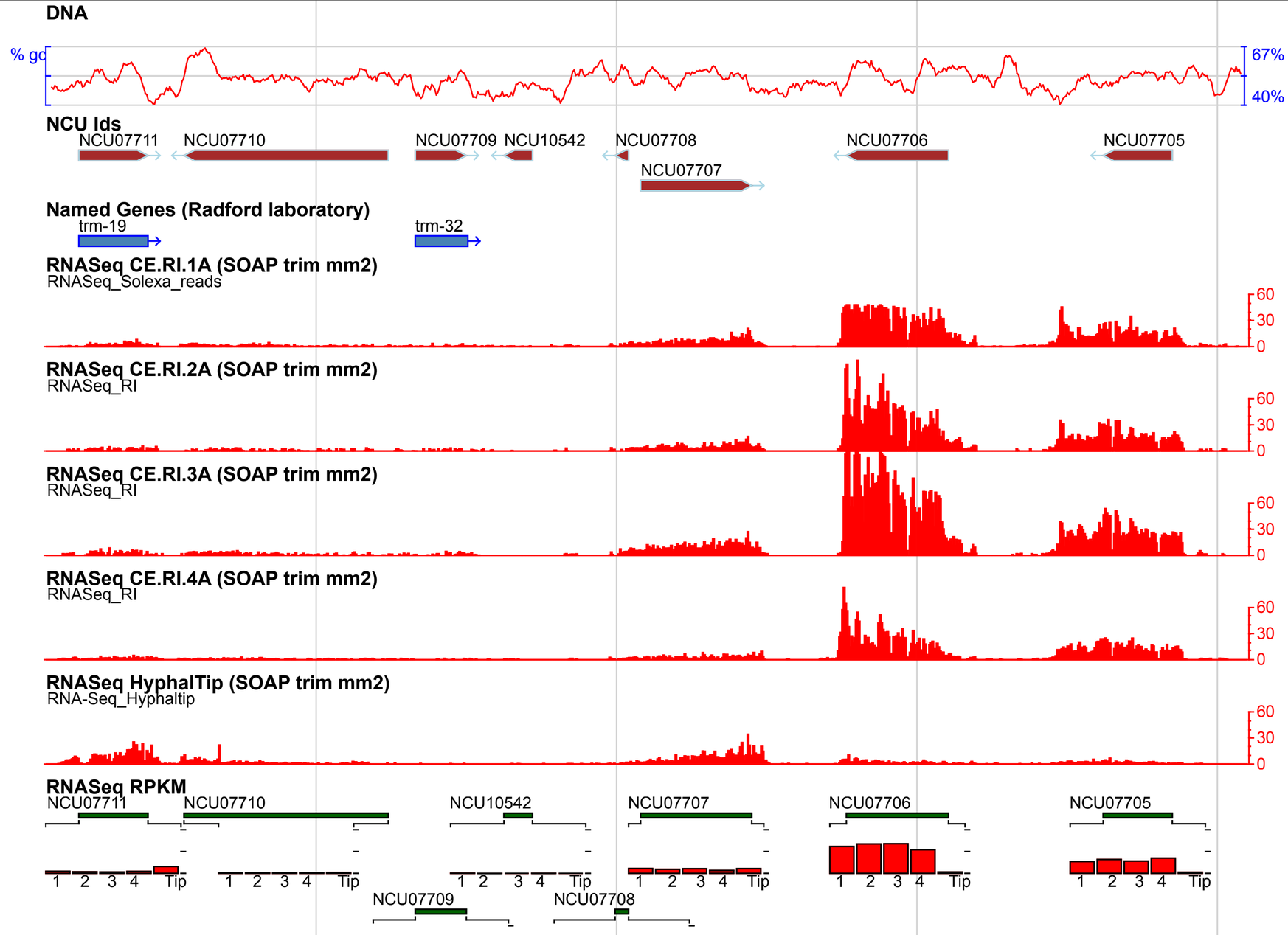
Estimating expression

- Calculating the average read-depth for a transcript
- Calculate the number of **R**EADS **P**ER **K**ILOBASE (of gene) per **M**illion Reads Mapped
- Normalized between libraries (Million Reads Mapped) and among genes in same genome (Per Kilobase) to determine an expression level for each gene in each library
- Plenty of caveats but it tends to produce reproducible results

RPKM is provides reliable indicator of concentration

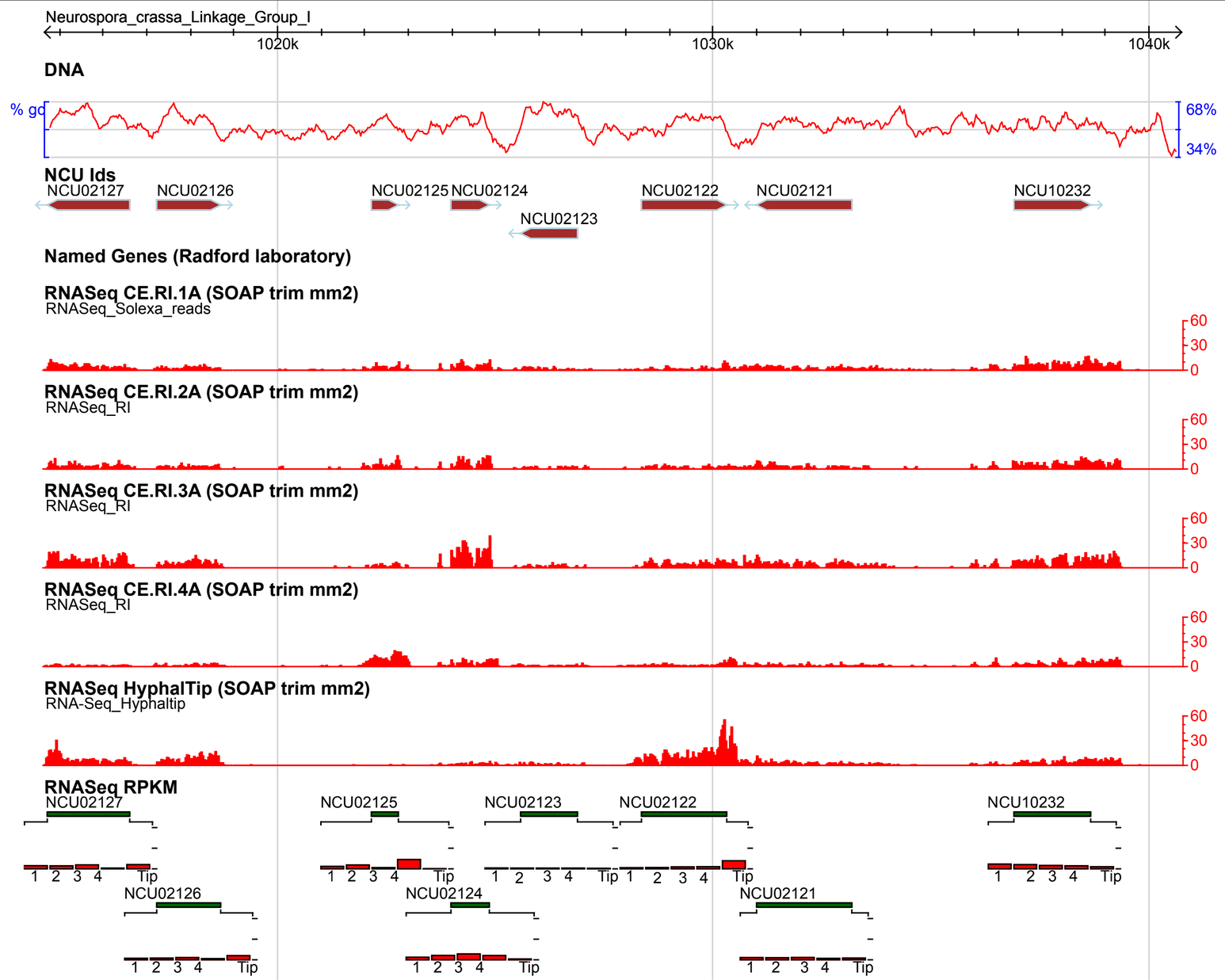


Mortazavi et al, Nat Methods 2008



RNA-Seq Raw and RPKM data for 5 conditions

Neurospora crassa



RNA-Seq Raw and RPKM data for 5 conditions

Neurospora crassa

Neurospora_crassa_Linkage_Group_III

2380.3k 2380.4k 2380.5k 2380.6k 2380.7k 2380.8k 2380.9k 2381k 2381.1k 2381.2k 2381.3k 2381.4k

DNA

% gc 100% 25%

RNASeq RPKM

NCU11171

1 2 3 4 Tip

NCU Ids

NCU11171

Named Genes (Radford laboratory)

mfa-1

GeneMark+PASA updated

3772_g

3772_g.T0

3772_g.T1

3772_g.T2

RNASeq CE.RI.3A (SOAP trim mm2)

RNASeq_RI

60 30 0

N.crassa PASA cDNA

All Neurospora ESTs PASA cDNA

asmbl_7178

asmbl_7180

asmbl_7179

asmbl_7181

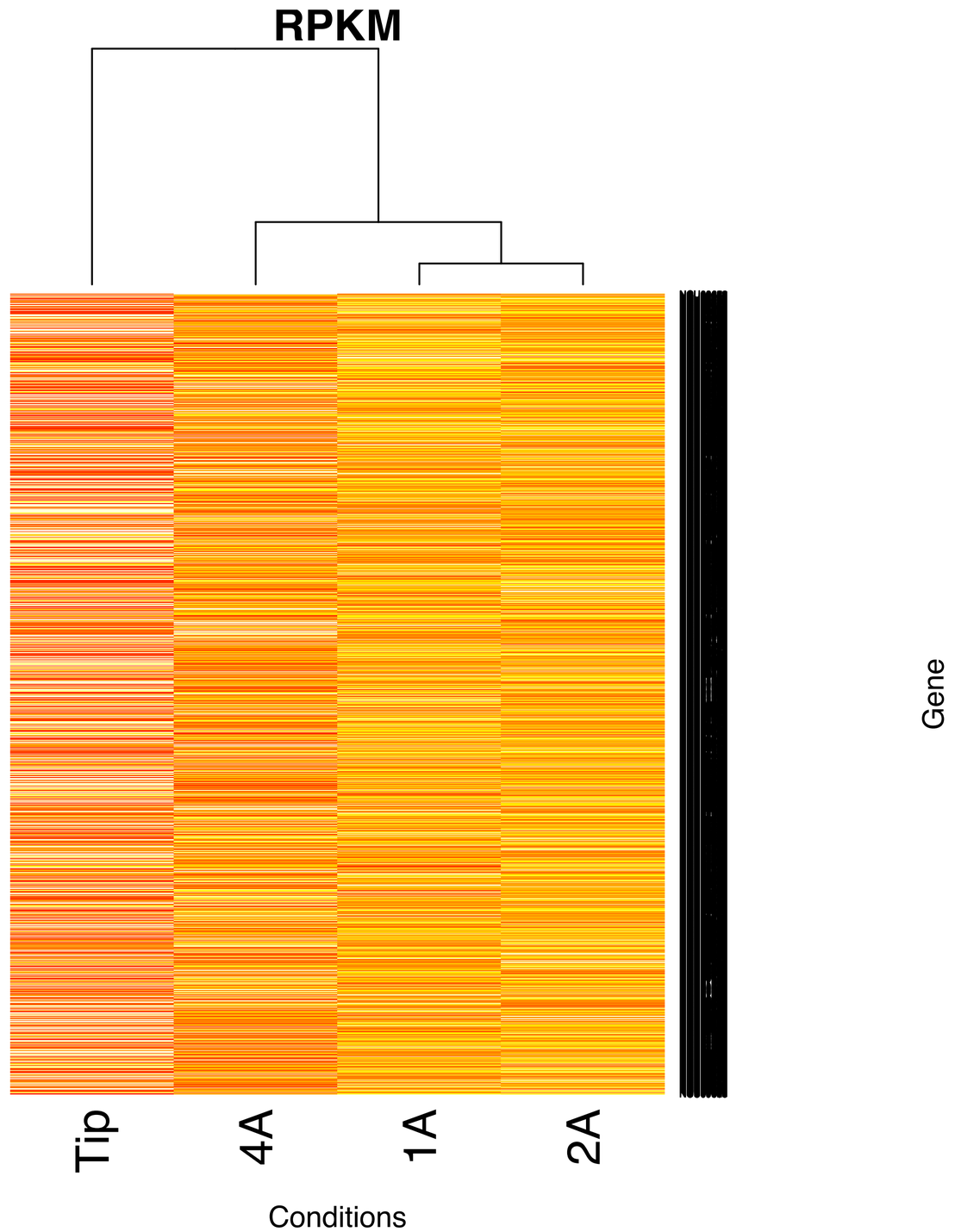
K4me2 ChIP-Seq

ChIPSeq_K4dime

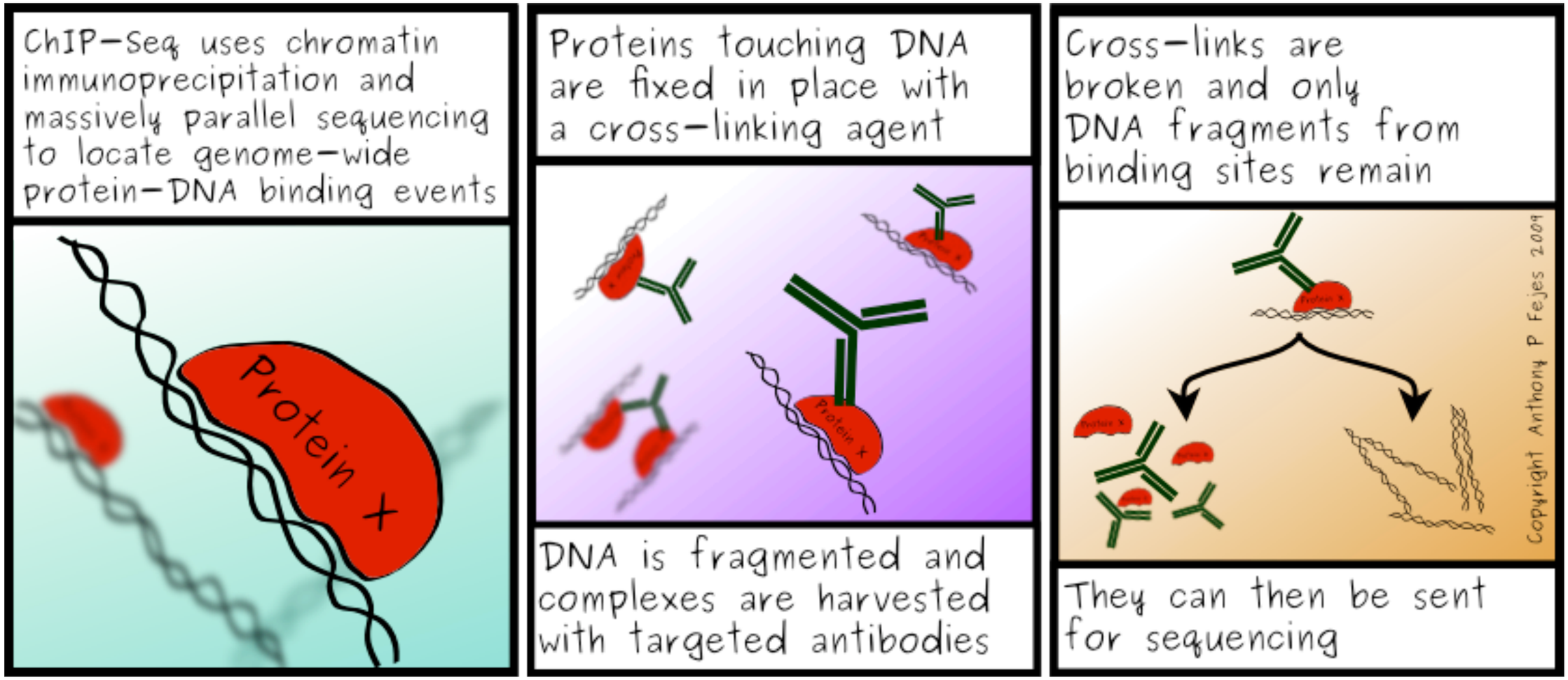
30 15 0

Clustering RPKM

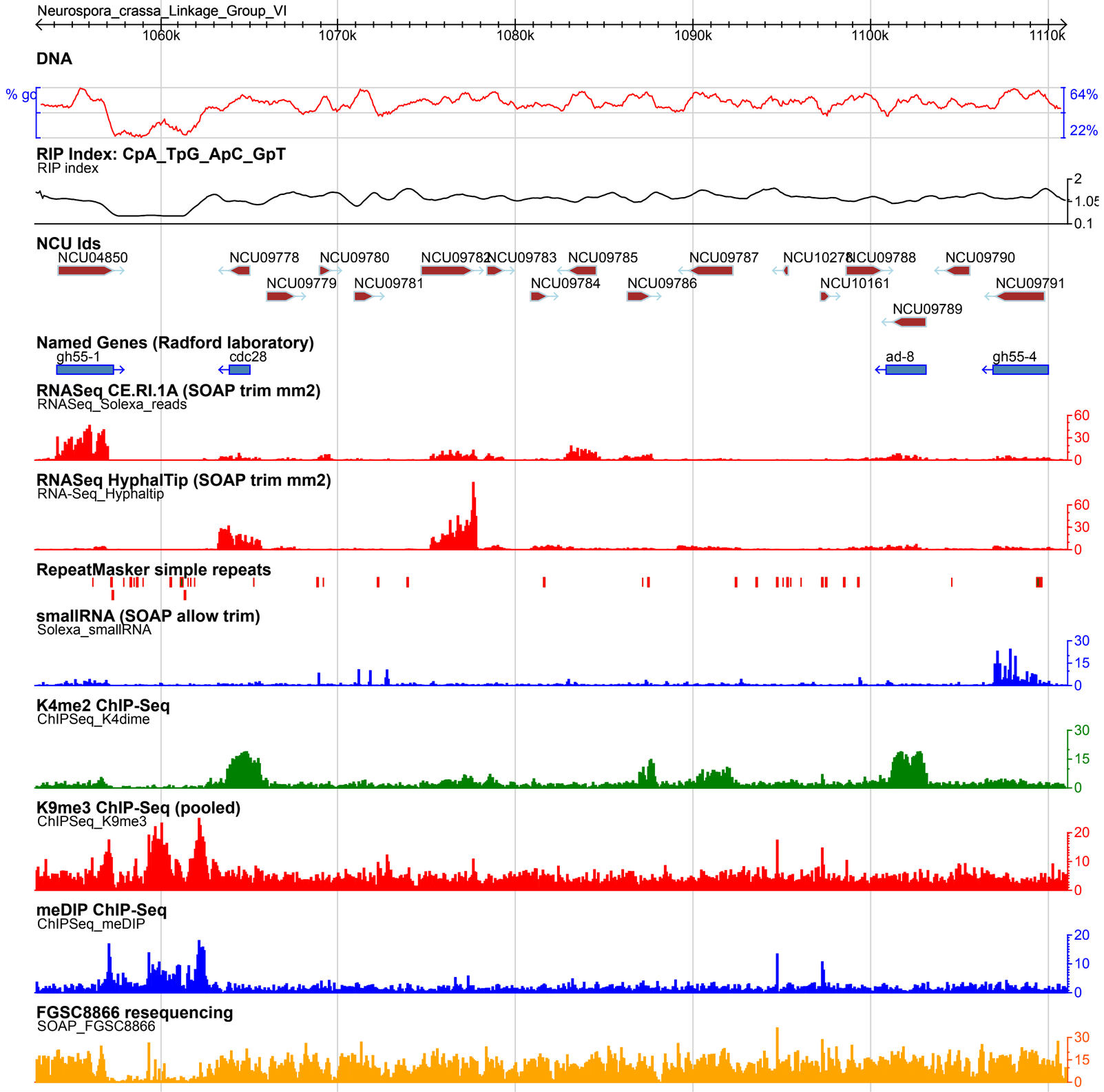
Can treat RPKM like expression values and cluster in hierarchal clustering like Microarray data



ChIP-Seq



Anthony Fejes - fejes.ca

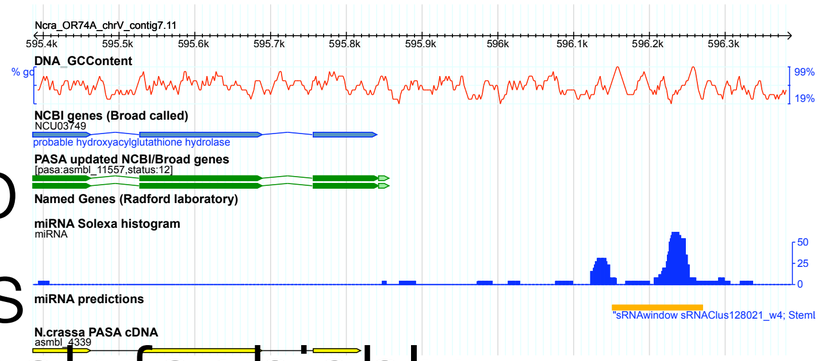

Neurospora_crassa_Linkage_Group_VI
1060k
1070k
1080k
1090k
1100k
1110k
DNA
% gc
64%
22%
RIP Index: CpA_TpG_ApC_GpT
RIP index
2
1.05
0.1
NCU Ids
NCU04850
NCU09778
NCU09780
NCU09782
NCU09783
NCU09785
NCU09787
NCU10278
NCU09788
NCU09790
NCU09779
NCU09781
NCU09784
NCU09786
NCU10161
NCU09791
NCU09789
Named Genes (Radford laboratory)
gh55-1
cdc28
ad-8
gh55-4
RNASeq CE.RI.1A (SOAP trim mm2)
RNASeq_Solexa_reads
60
30
0
RNASeq HyphalTip (SOAP trim mm2)
RNA-Seq_Hyphaltip
60
30
0
RepeatMasker simple repeats
smallRNA (SOAP allow trim)
Solexa_smallRNA
30
15
0
K4me2 ChIP-Seq
ChIPSeq_K4dime
30
15
0
K9me3 ChIP-Seq (pooled)
ChIPSeq_K9me3
20
10
0
meDIP ChIP-Seq
ChIPSeq_meDIP
20
10
0
FGSC8866 resequencing
SOAP_FGSC8866
30
15
0

small RNA Sequencing

Map to Genome

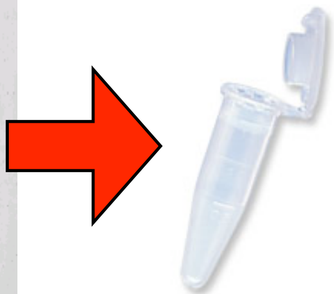
Extract RNA

~5M 36bp sequences



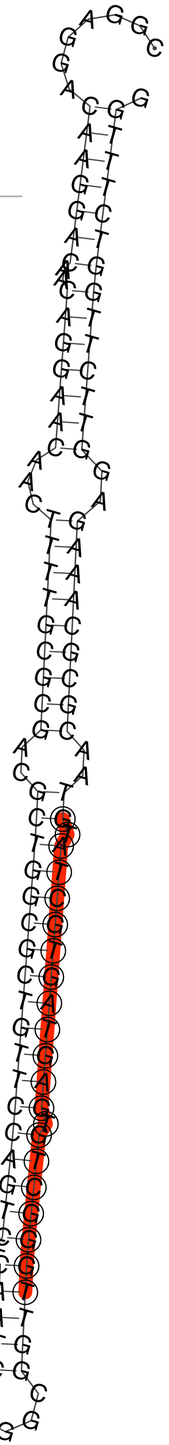
Solexa (Illumina) Sequencing

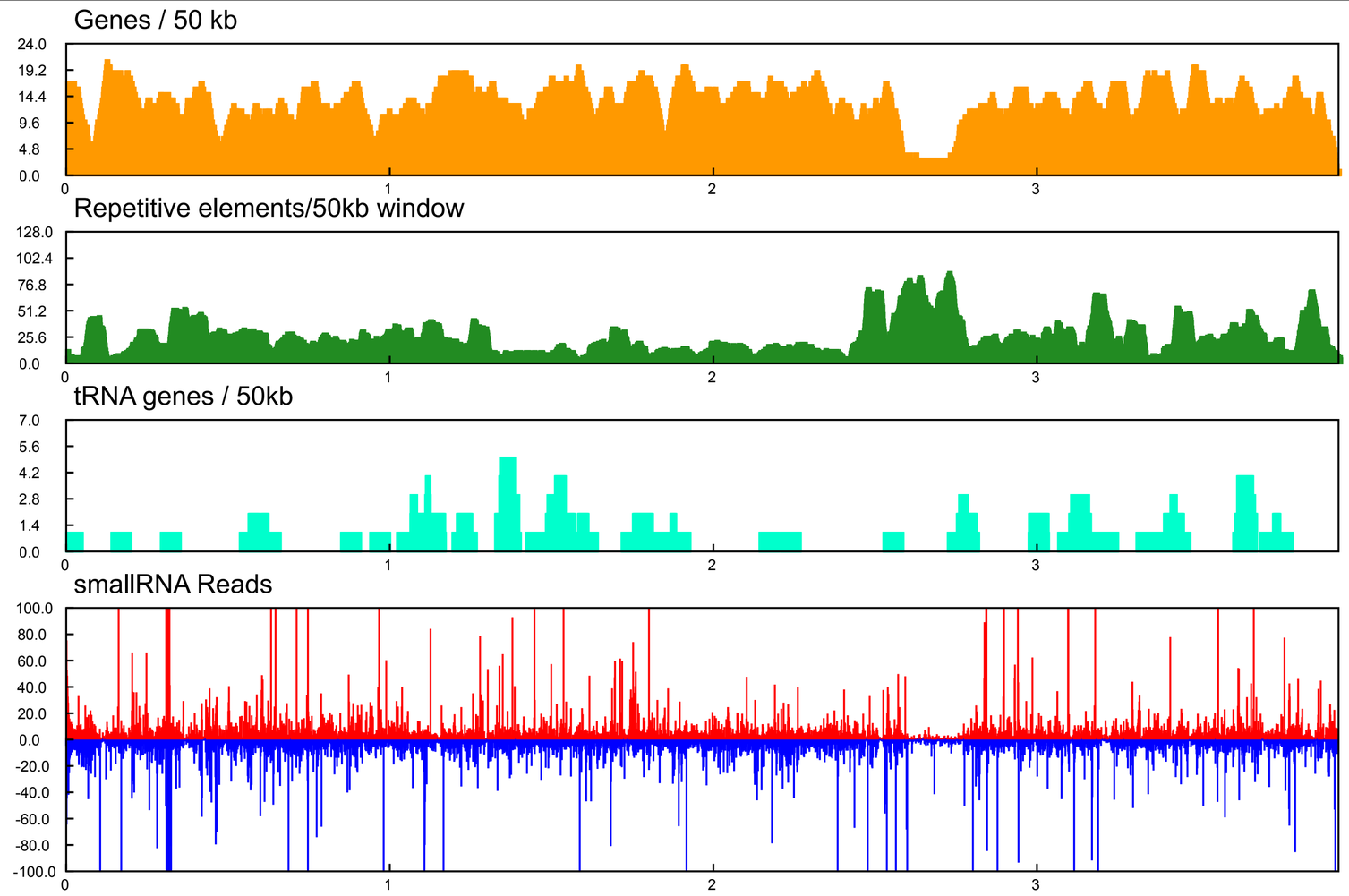
1. Look for highly expressed
2. Identify conserved secondary structure

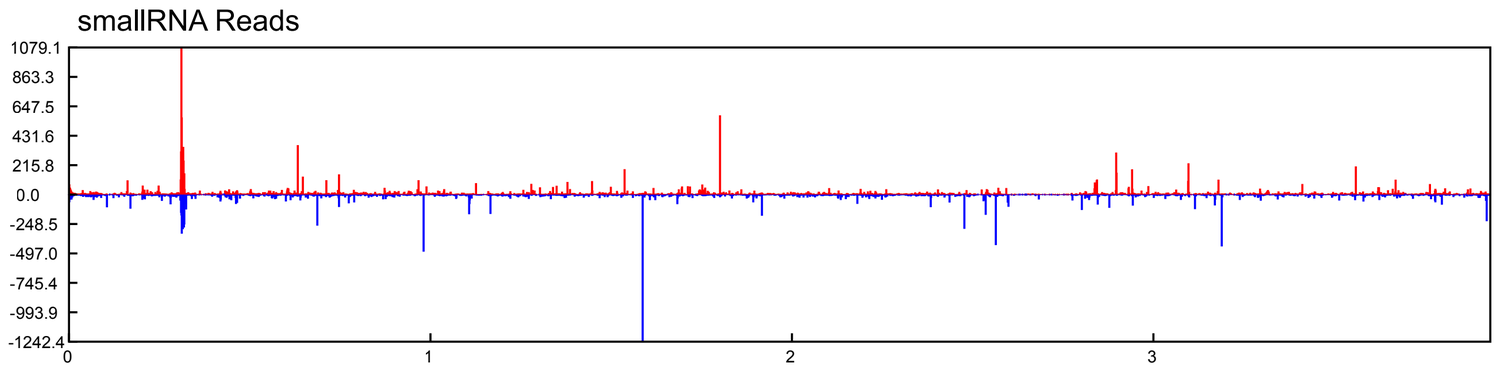


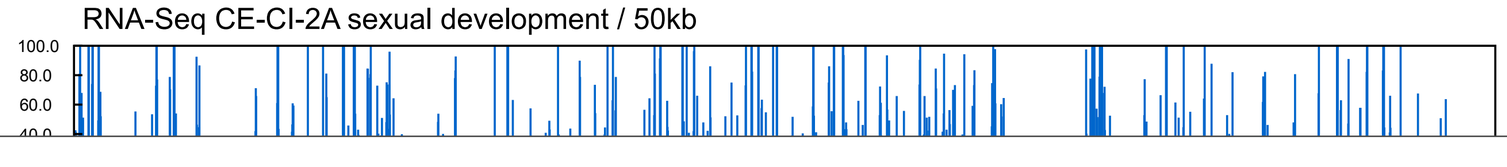
RNA cloning protocol

```
>n_crassa
CACGUGGGAUCGGGCACCCAUAAAGGGUCCGGACCCCCCGUCGUGGGCCAAAGCGGGGAACG
(((((((..(((((((.((......)))))))).))(((((..((...)).)))))).)))
>n_tetrasperma_2508
CACGUGGGAUCGGGCACCCAUAAAGGGUCCGGACCCCCCGUCGUGGGCCAAAGCGGGGAACG
(((((((..(((((((.((......)))))))).))(((((..((...)).)))))).)))
>n_discreta_8579
CACGUGGGAUCGGGCGCCCAAAAAAGGUCCGGGUCCCCCGUCGUGGGCCAAAGCGGGGAACG
((((.((((..((....)).....((.((((.((.((((((..((...)).)))))).)).
>consensus
CACGUGGGAUCGGGCACCCAUAAAGGGUCCGGACCCCCCGUCGUGGGCCAAAGCGGGGAACG
((((.(((.((.((.(((((.....))))))..)).))...))))...((..((((.(.(((
```

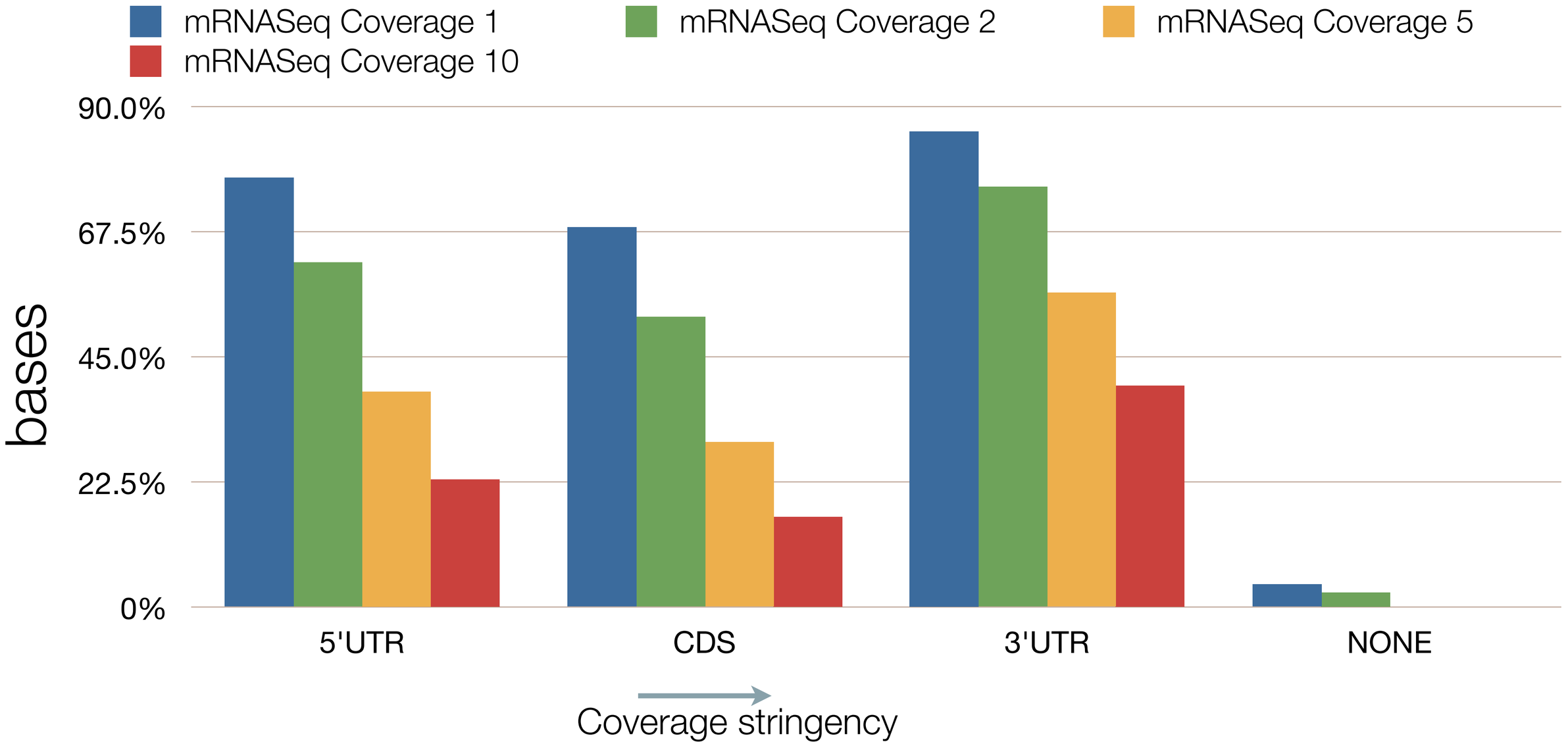



Genes / 50 kb
24.0
19.2
14.4
9.6
4.8
0.0
0
1
2
3
Repetitive elements/50kb window
128.0
102.4
76.8
51.2
25.6
0.0
0
1
2
3
tRNA genes / 50kb
7.0
5.6
4.2
2.8
1.4
0.0
0
1
2
3
smallRNA Reads
100.0
80.0
60.0
40.0
20.0
0.0
-20.0
-40.0
-60.0
-80.0
-100.0
0
1
2
3

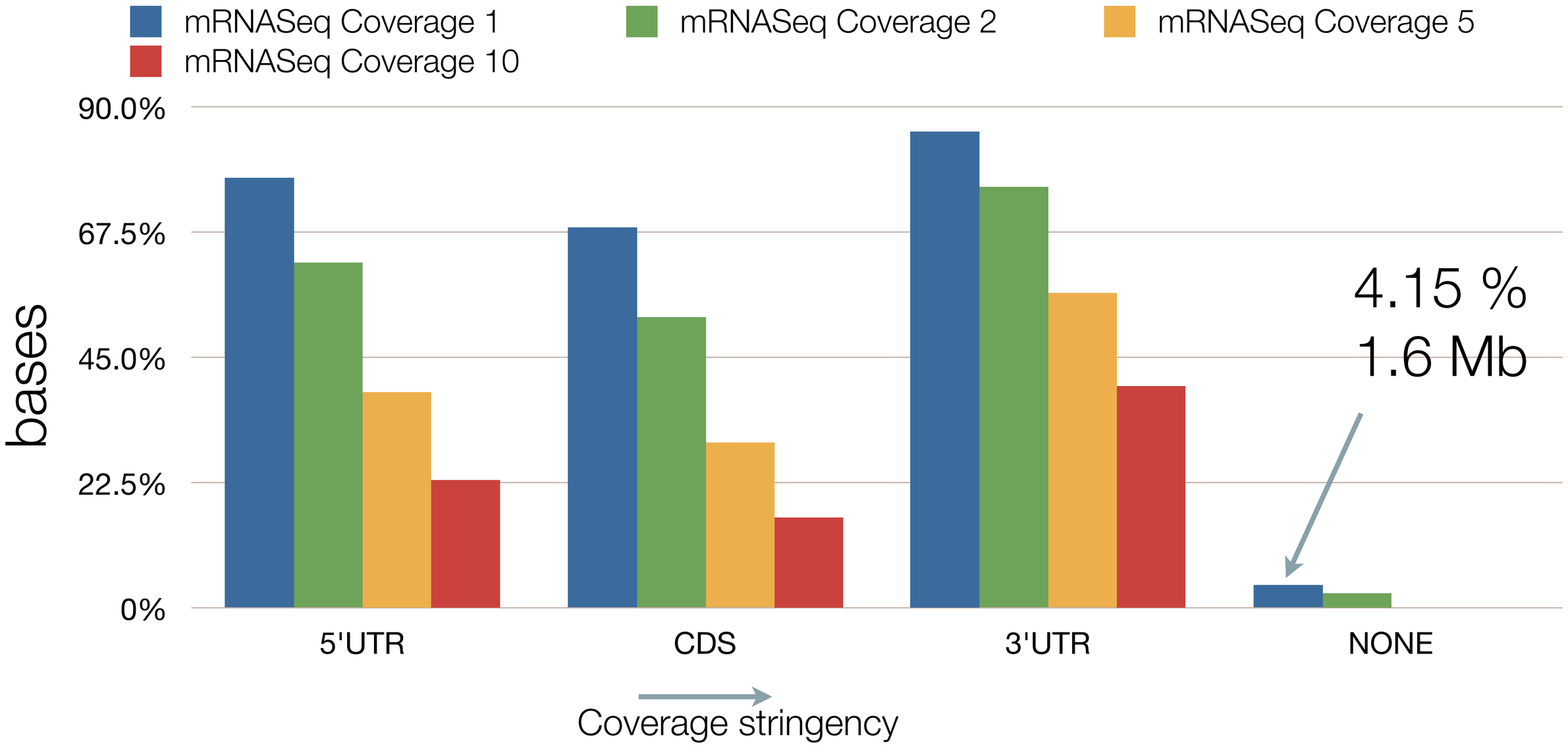

smallRNA Reads
1079.1
863.3
647.5
431.6
215.8
0.0
-248.5
-497.0
-745.4
-993.9
-1242.4
0
1
2
3


RNA-Seq CE-CI-2A sexual development / 50kb
100.0
80.0
60.0

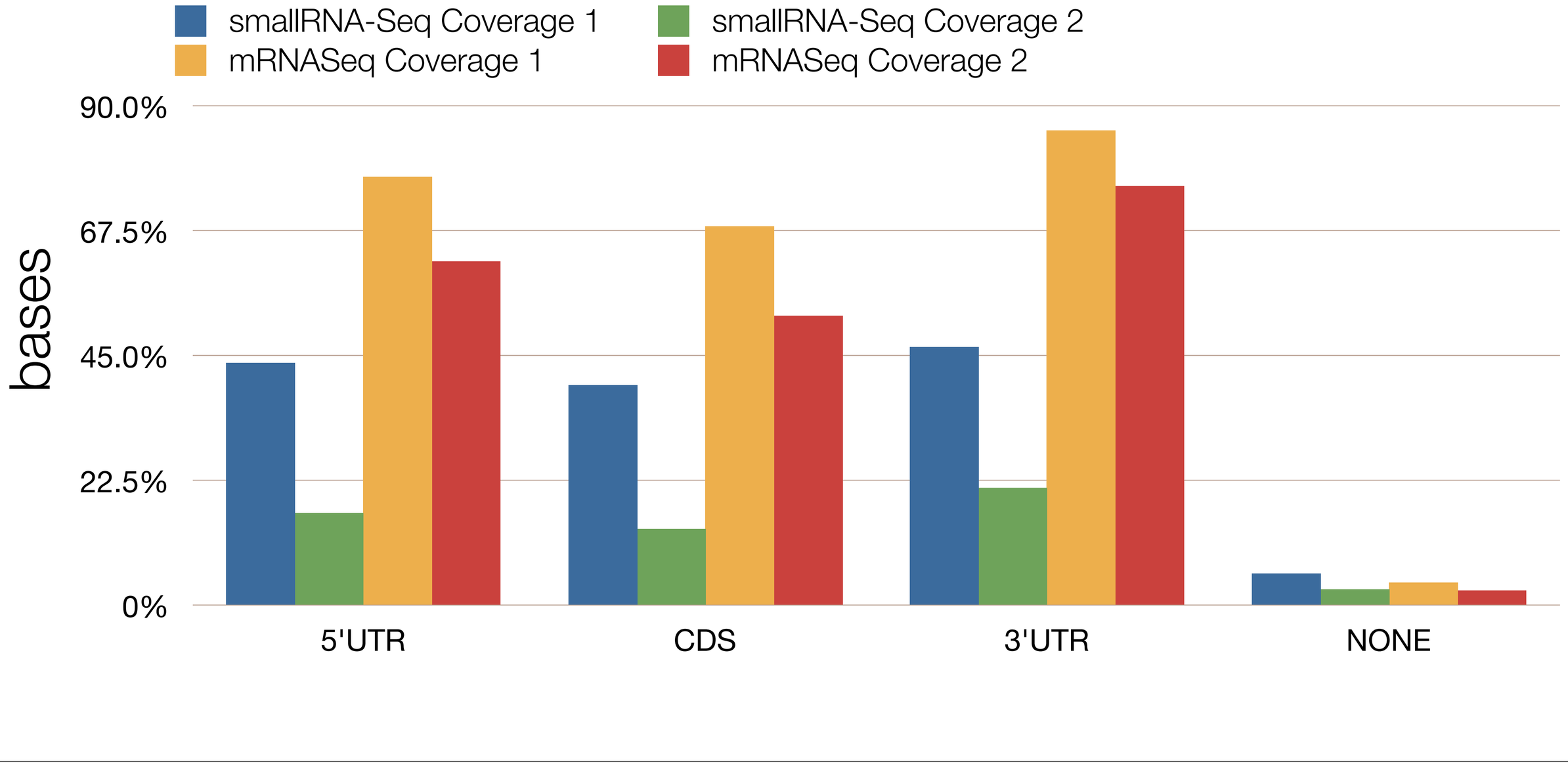
mRNASeq coverage of gene regions



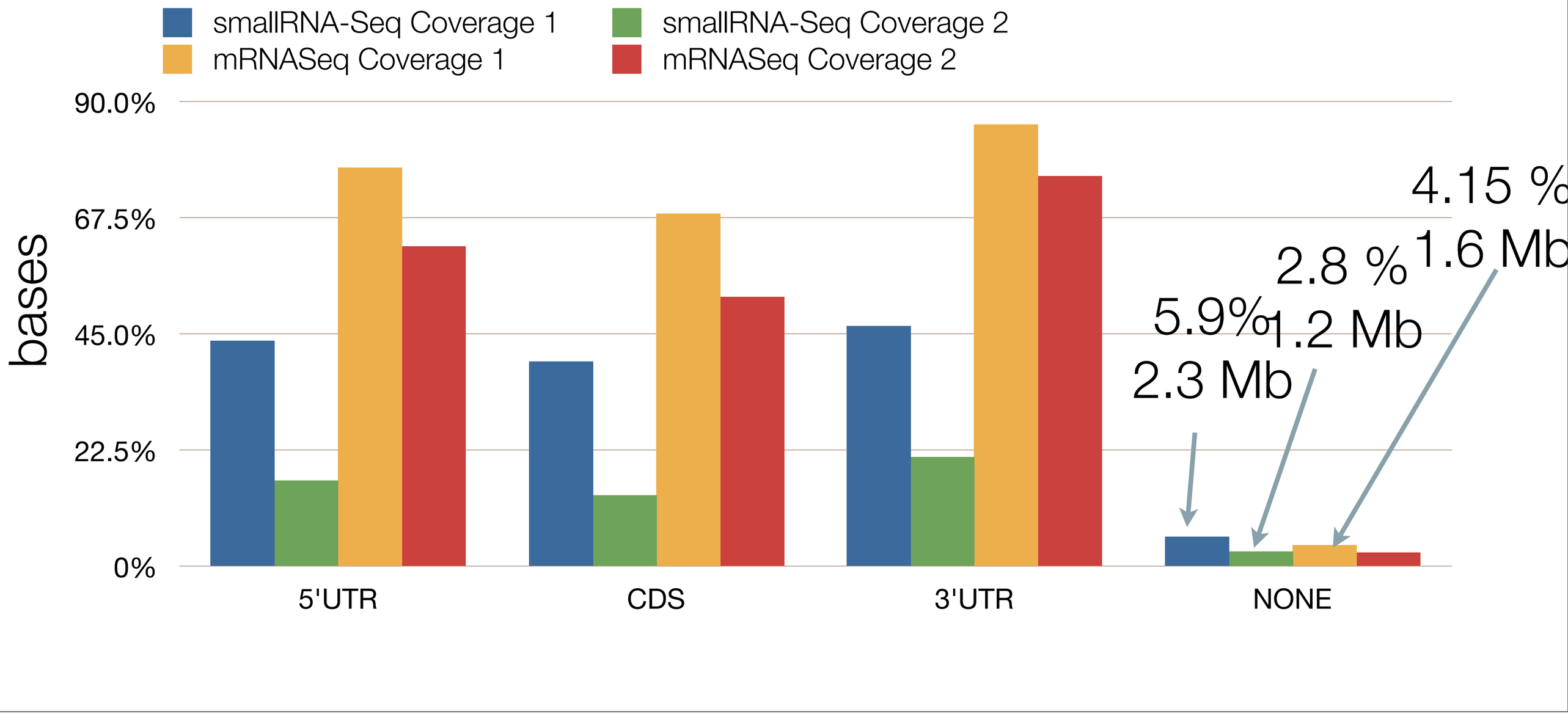
mRNASeq coverage of gene regions



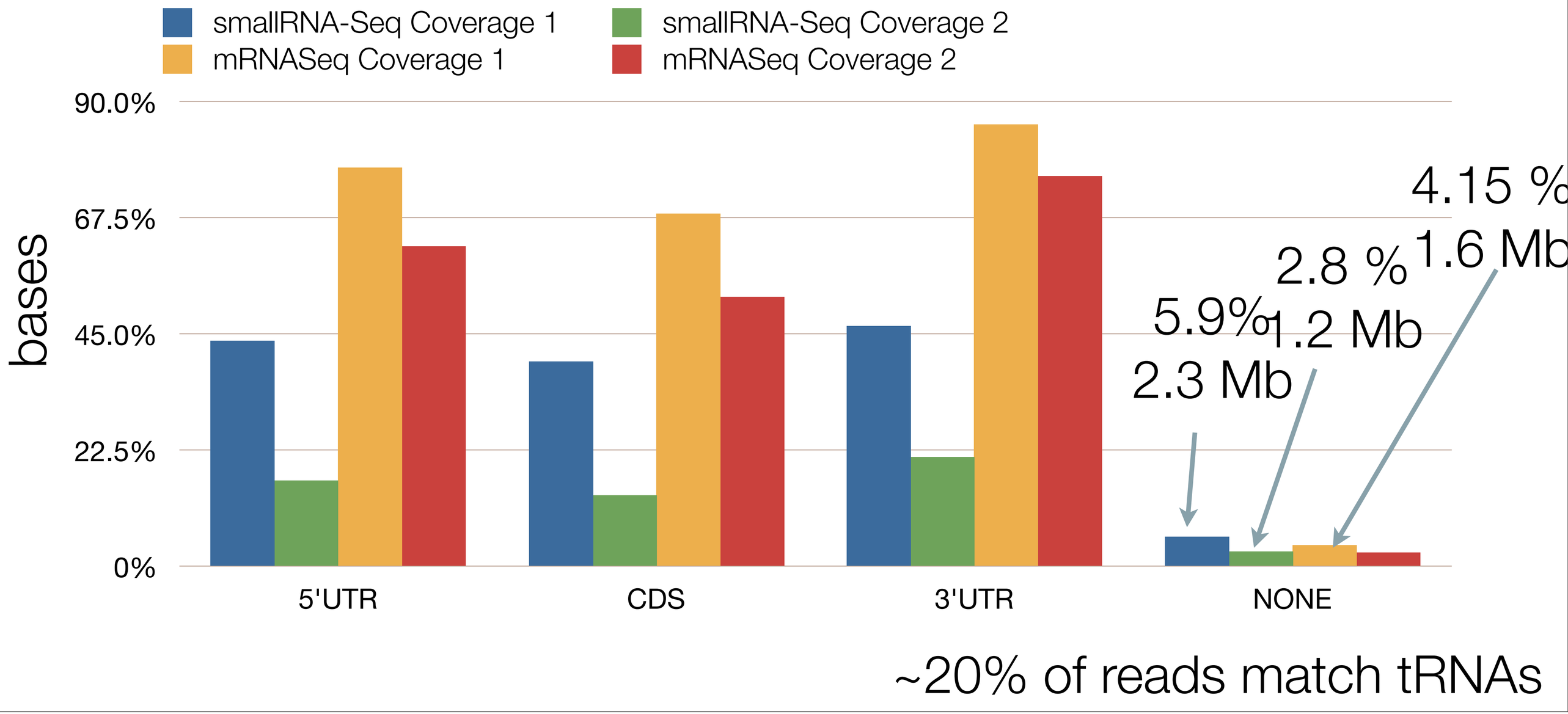
SmallRNA seq also covers lots of genic regions



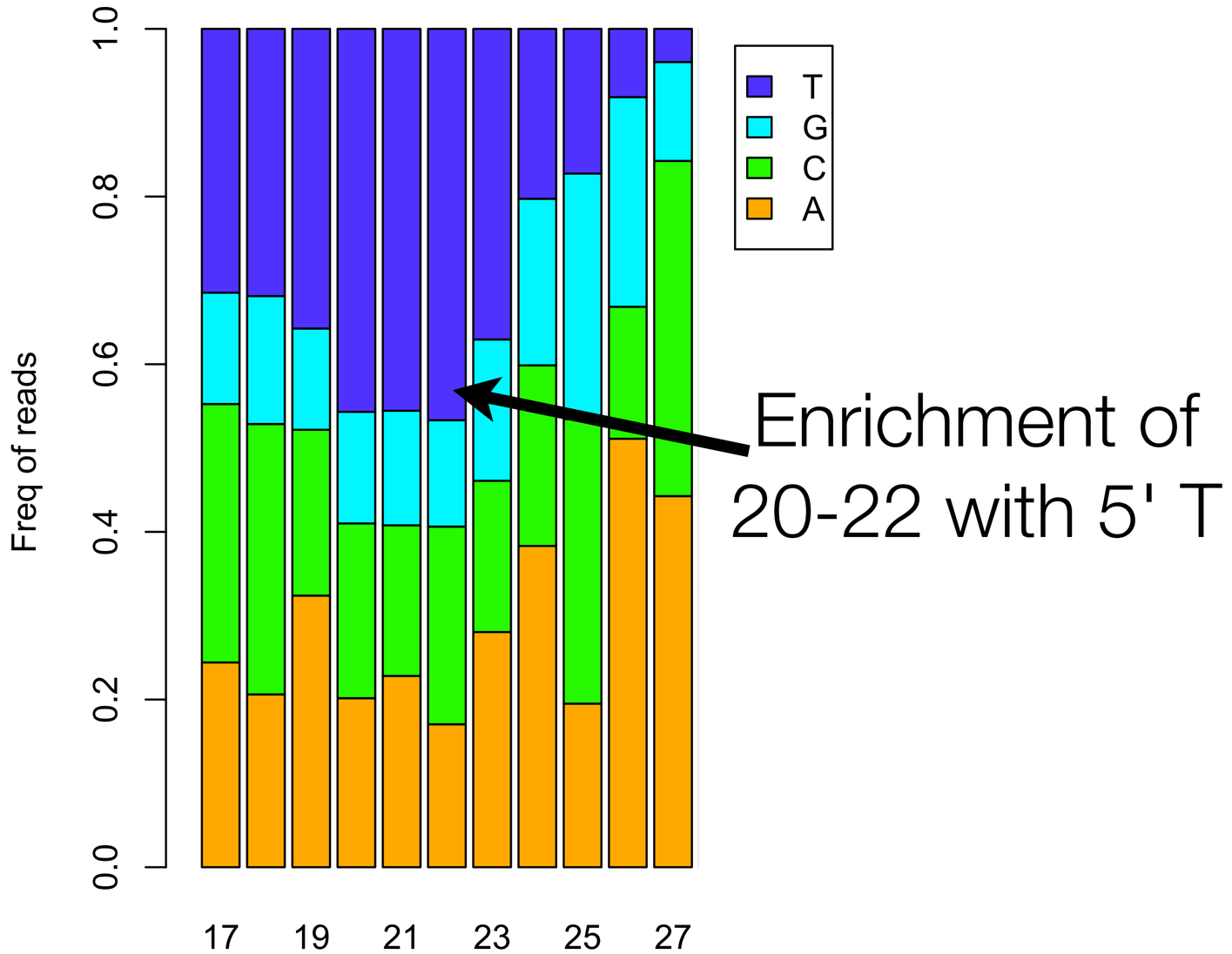
SmallRNA seq also covers lots of genic regions



SmallRNA seq also covers lots of genic regions



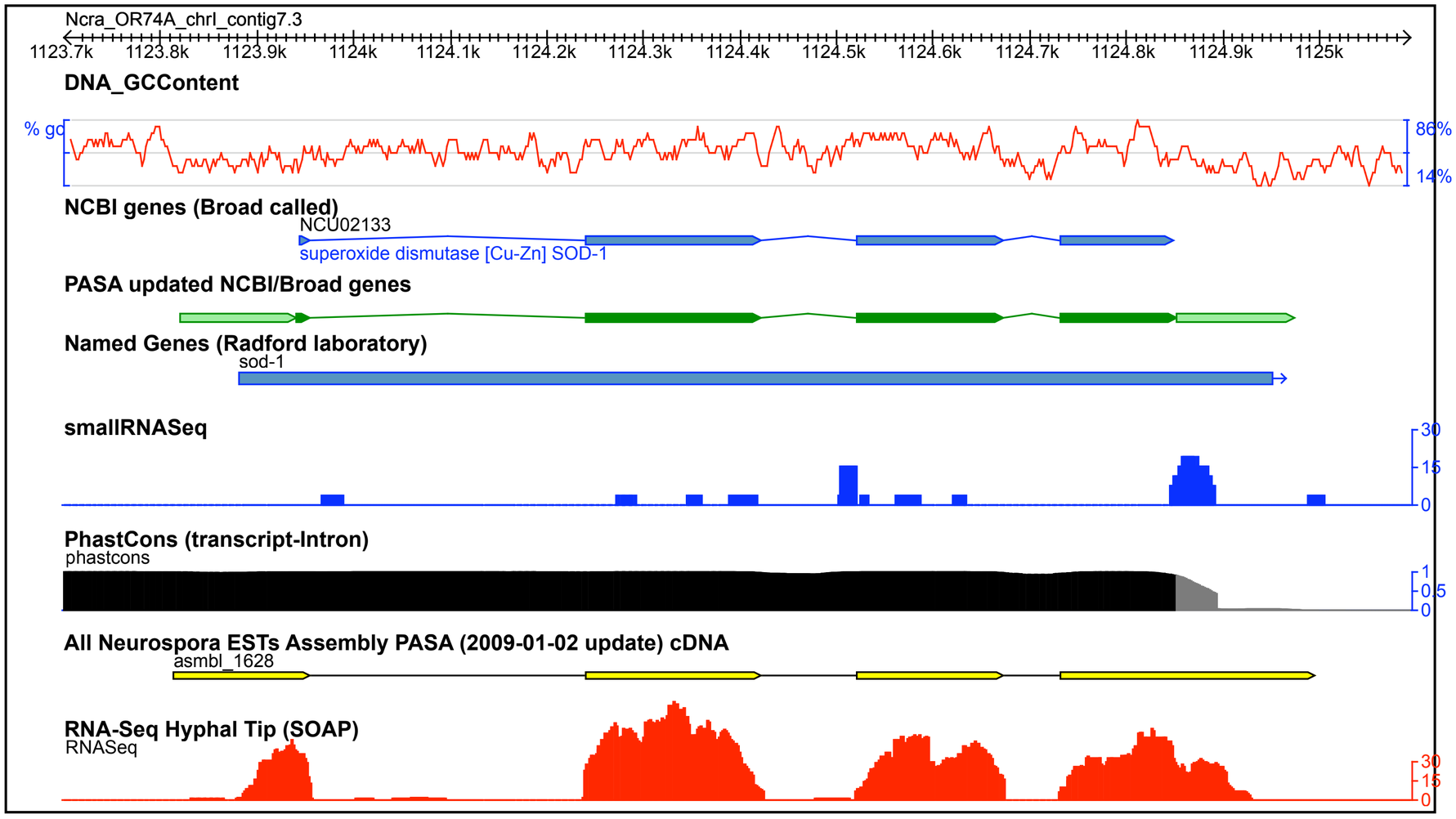
Size and sequence bias of sequenced smallRNA reads indicative of Dicer processing



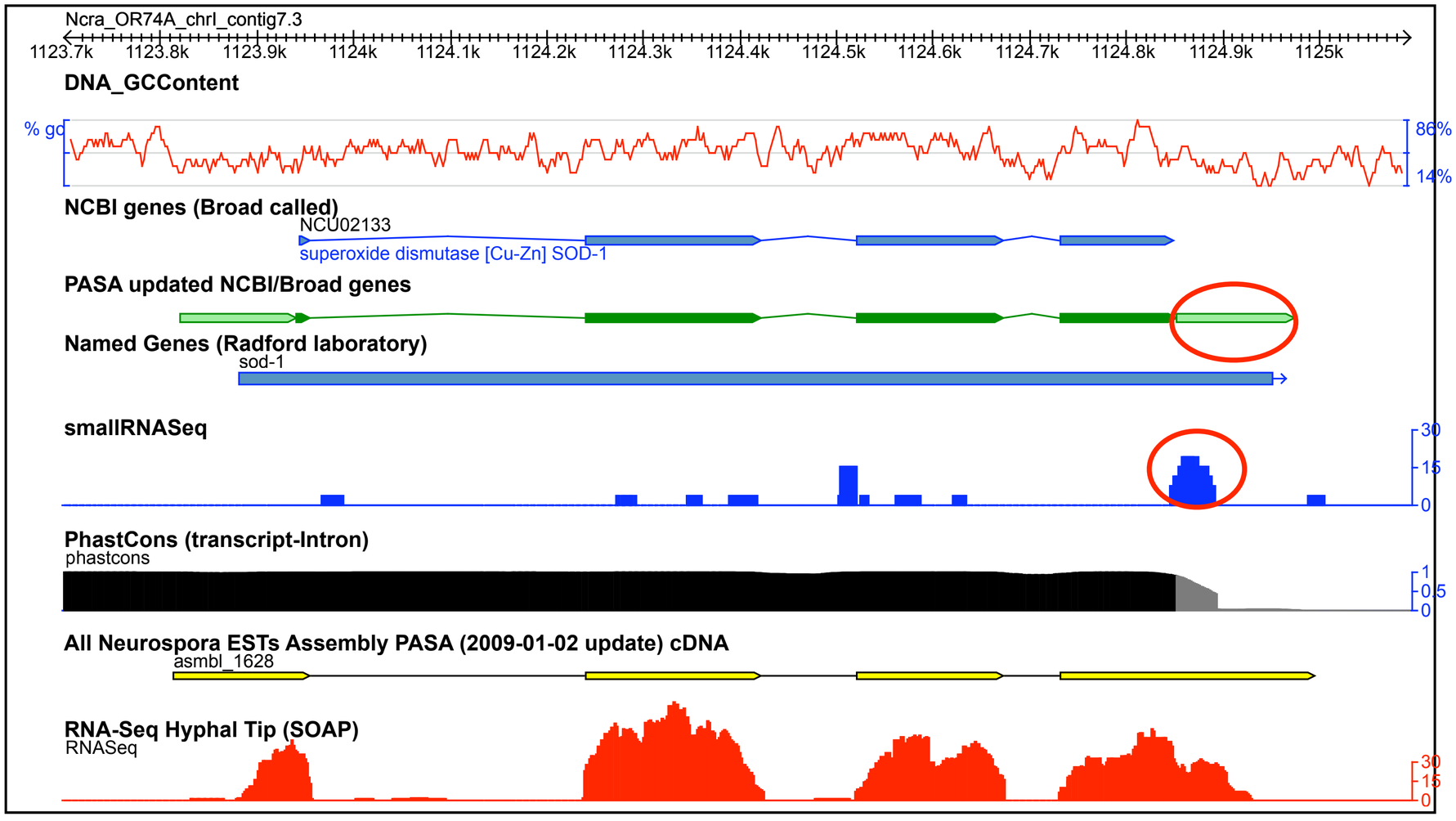
Putative Noncoding RNAs

- Several computational and comparative approaches
    - Hard to train models because there are no positive controls
- Prioritize candidates for noncoding RNAs discovery on
    - Highly expressed
    - Intergenic (not tRNA, rDNA)
    - UTR related

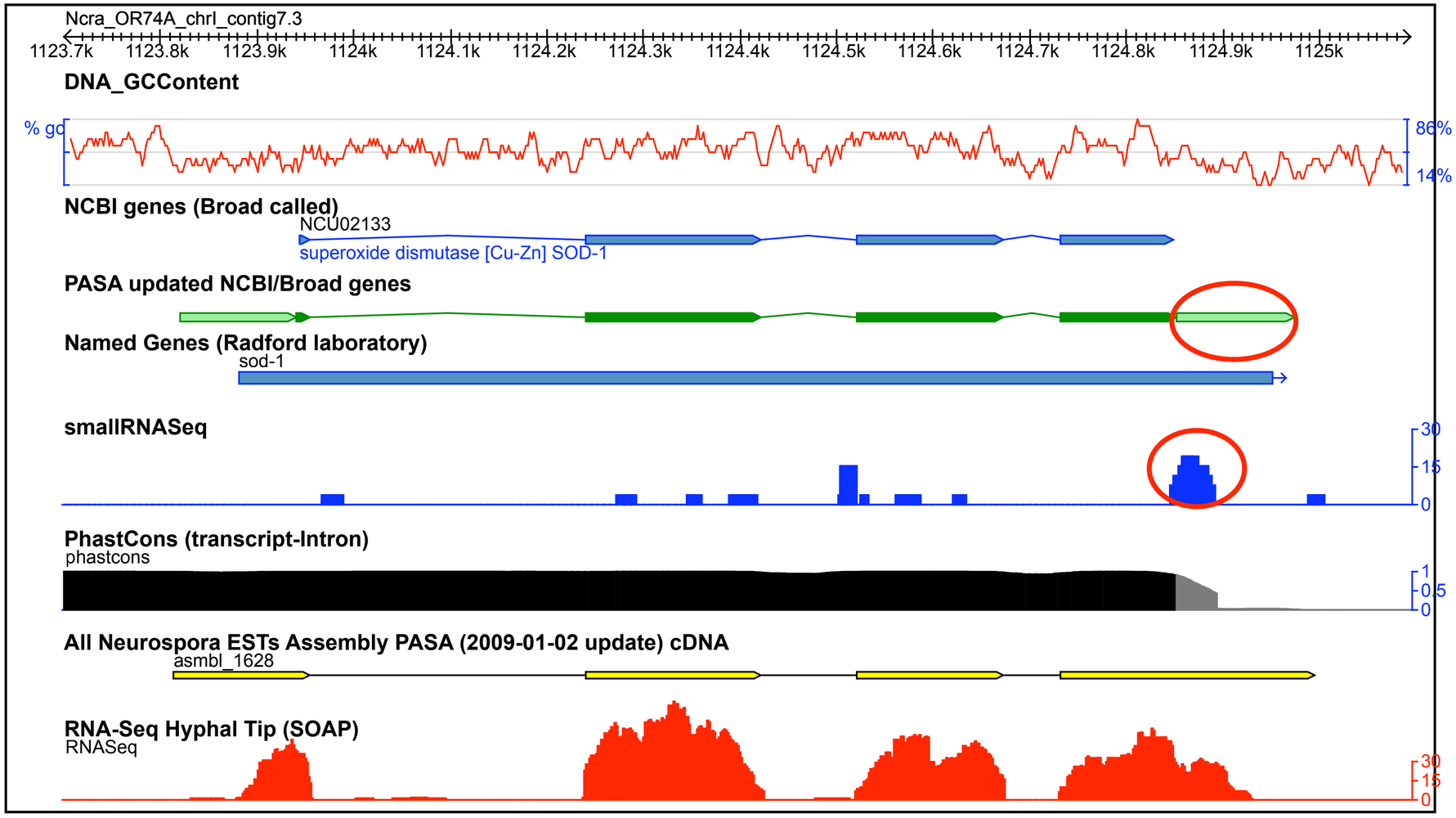
3' UTR, small RNAs, and Folding

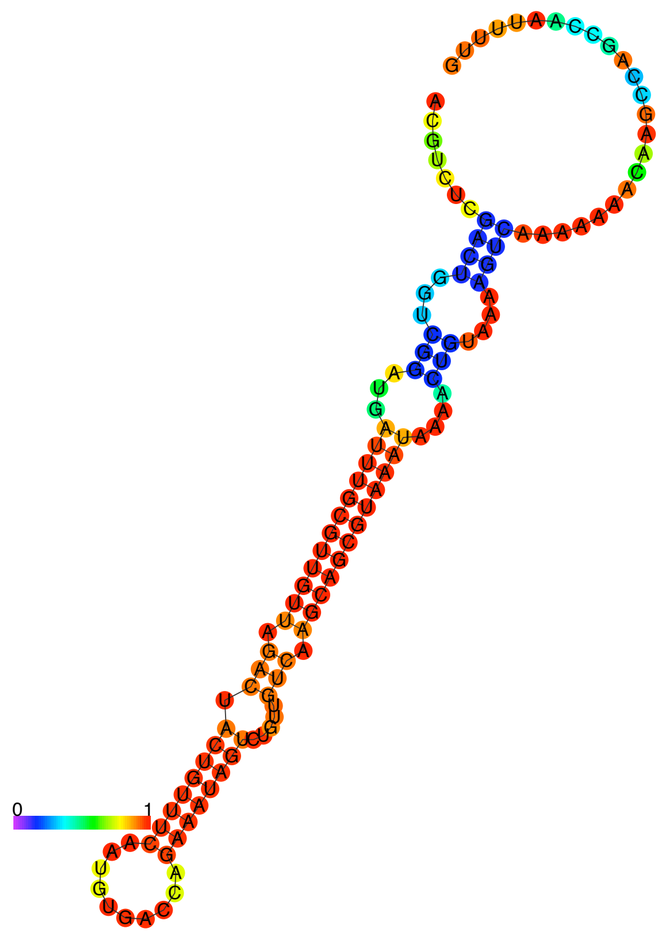


3' UTR, small RNAs, and Folding

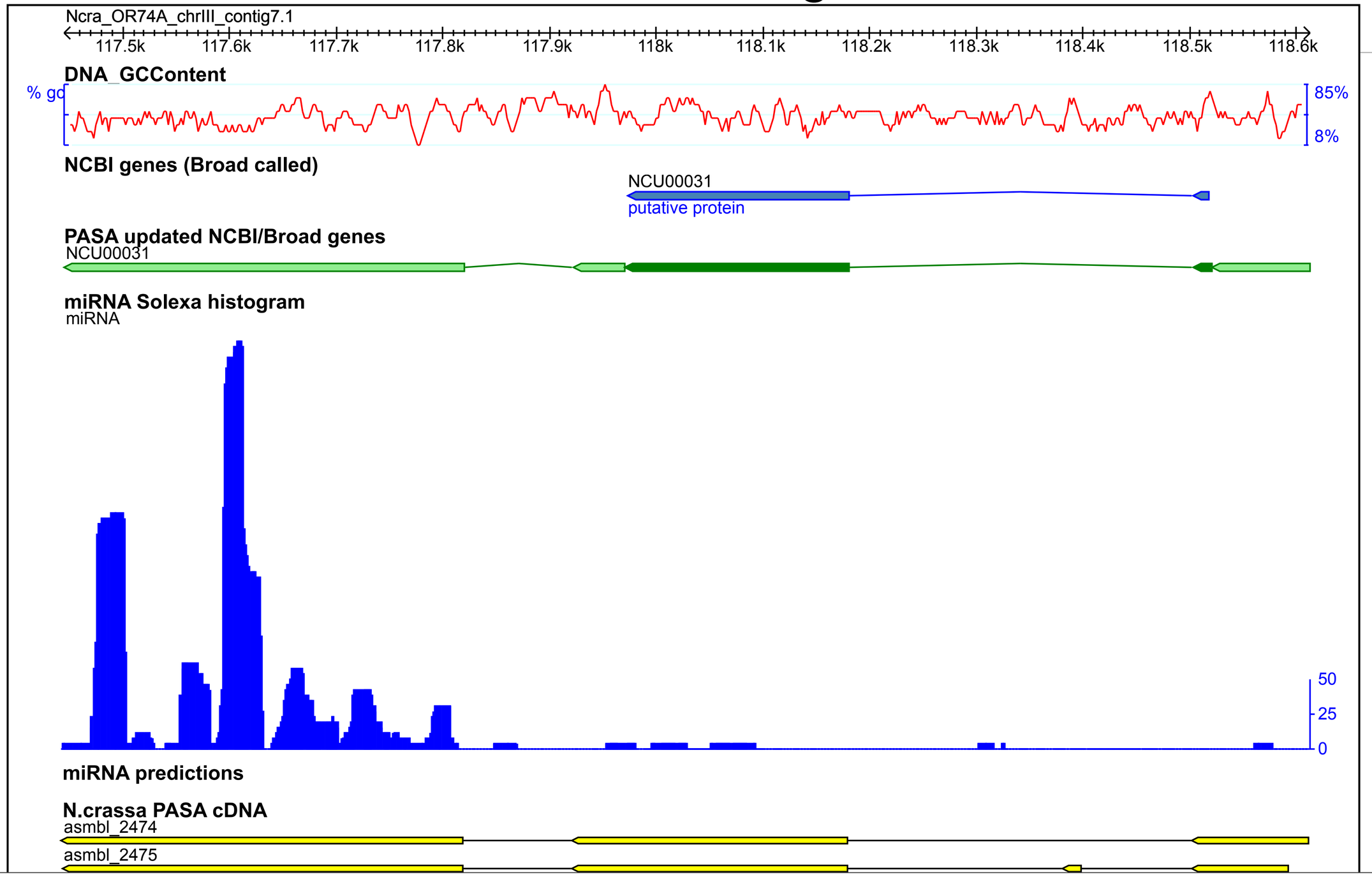


3' UTR, small RNAs, and Folding

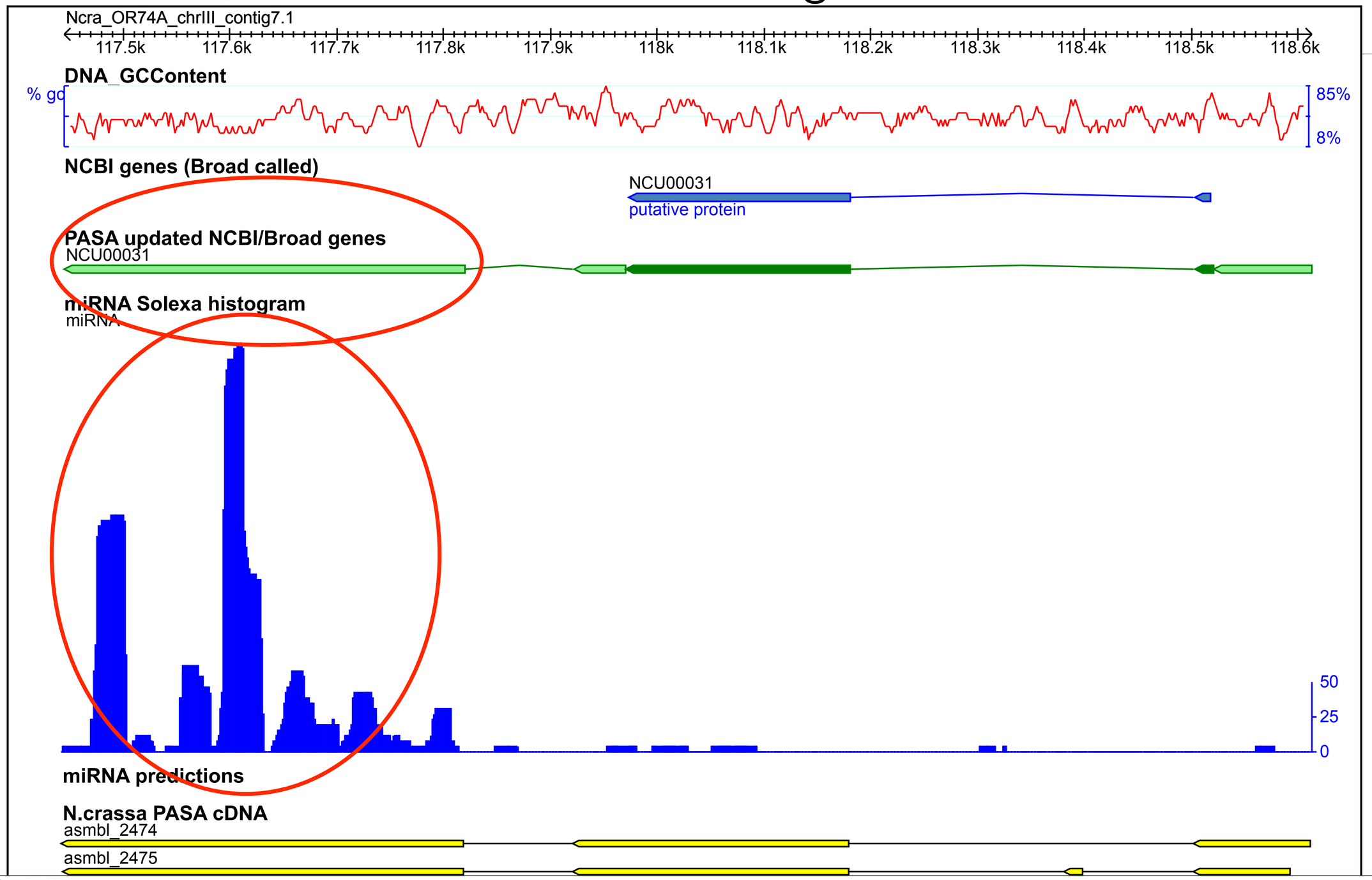




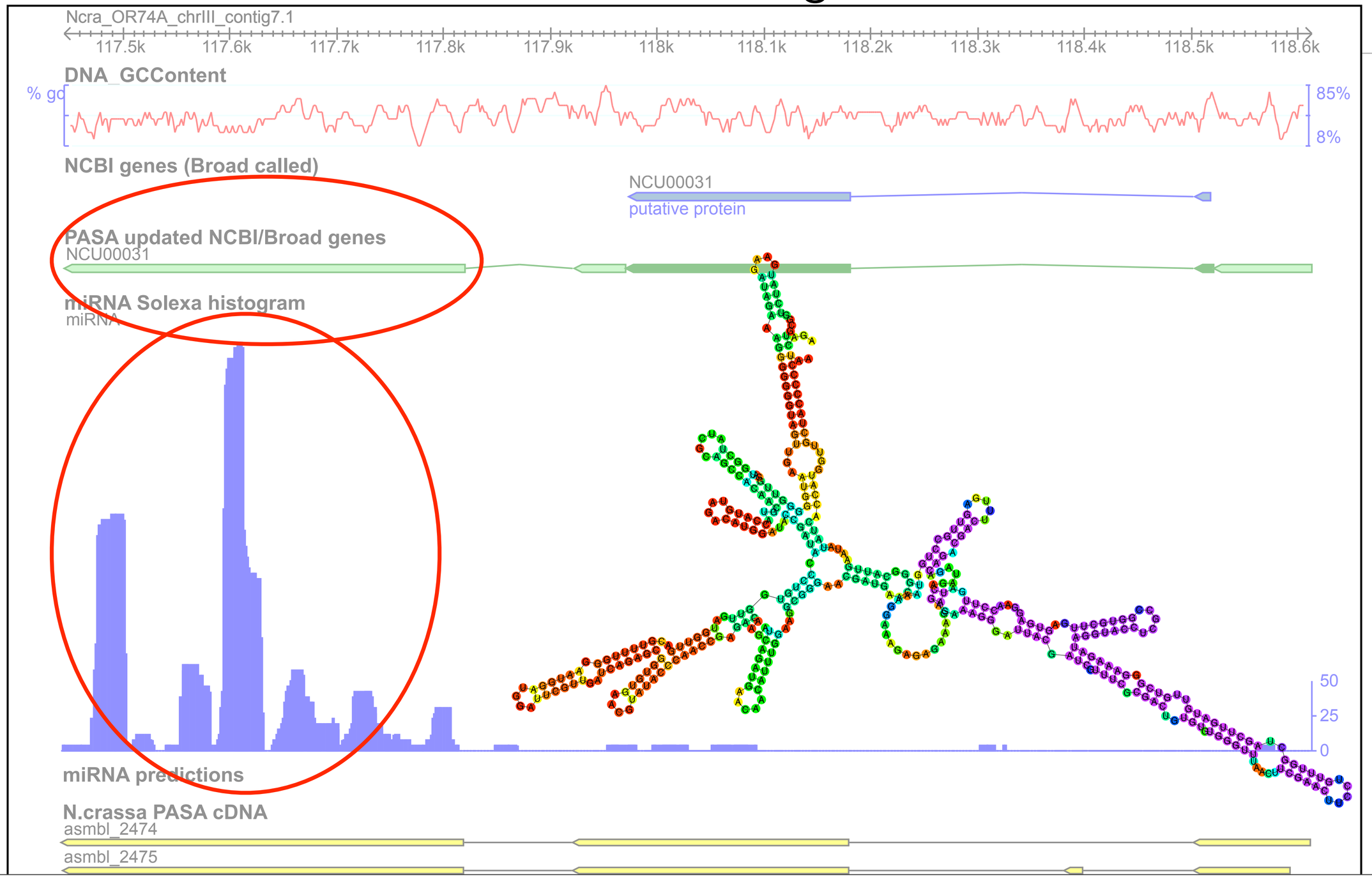
3' UTR, small RNAs, and Folding



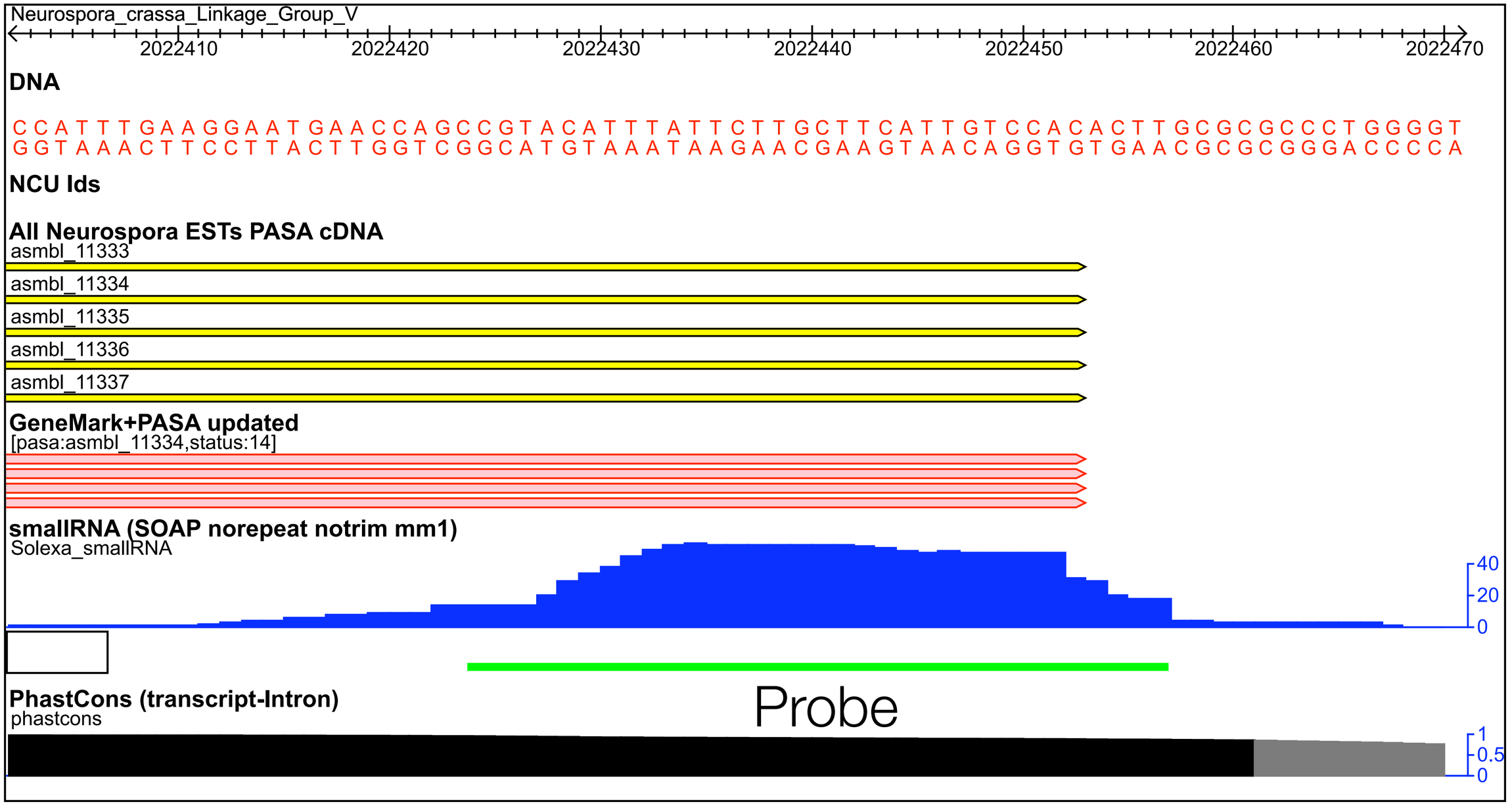
3' UTR, small RNAs, and Folding



3' UTR, small RNAs, and Folding

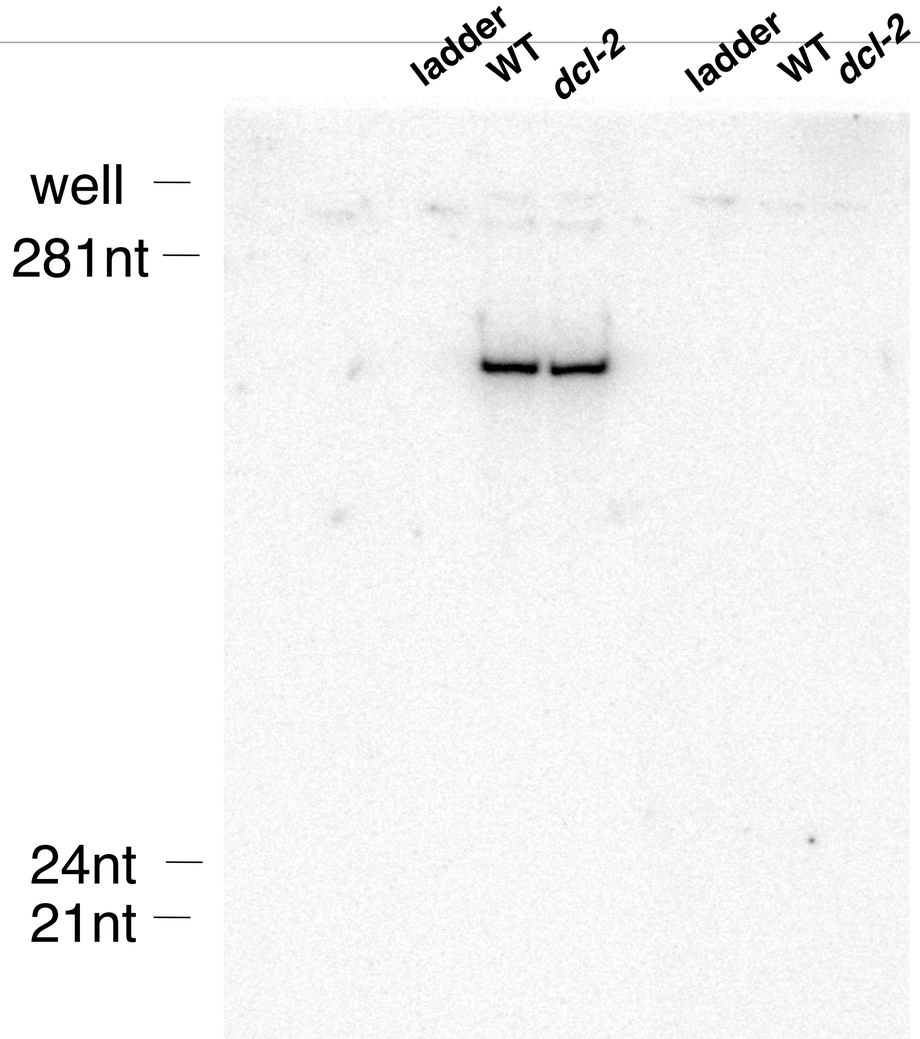


Candidate region



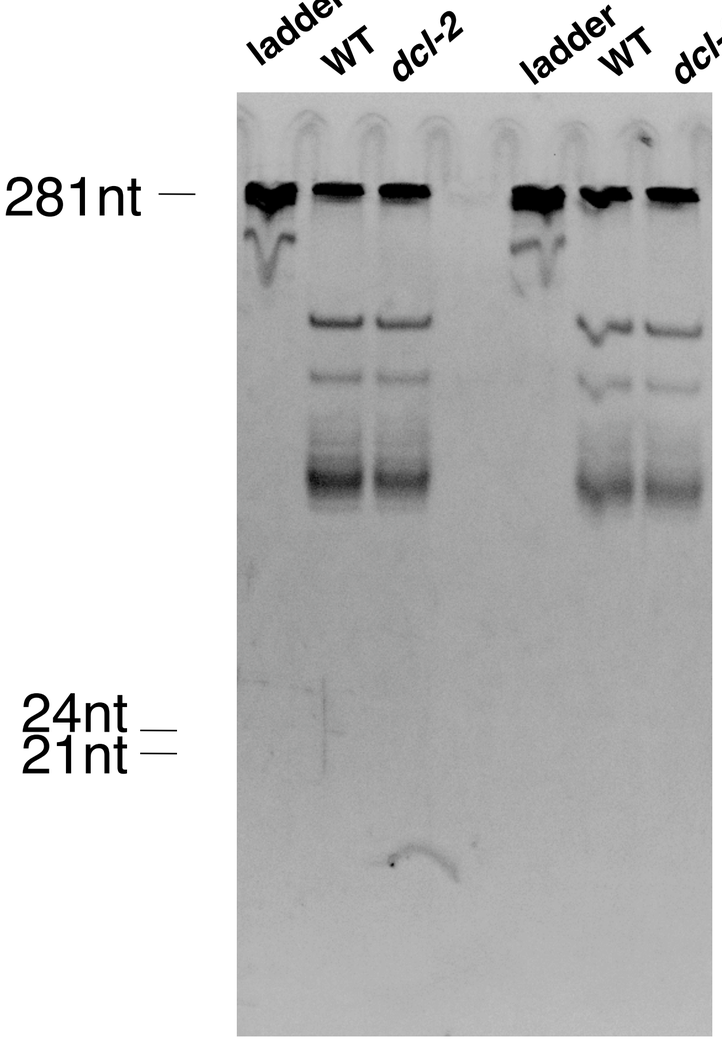
Kristina Smith

Confirmation by Northern blots



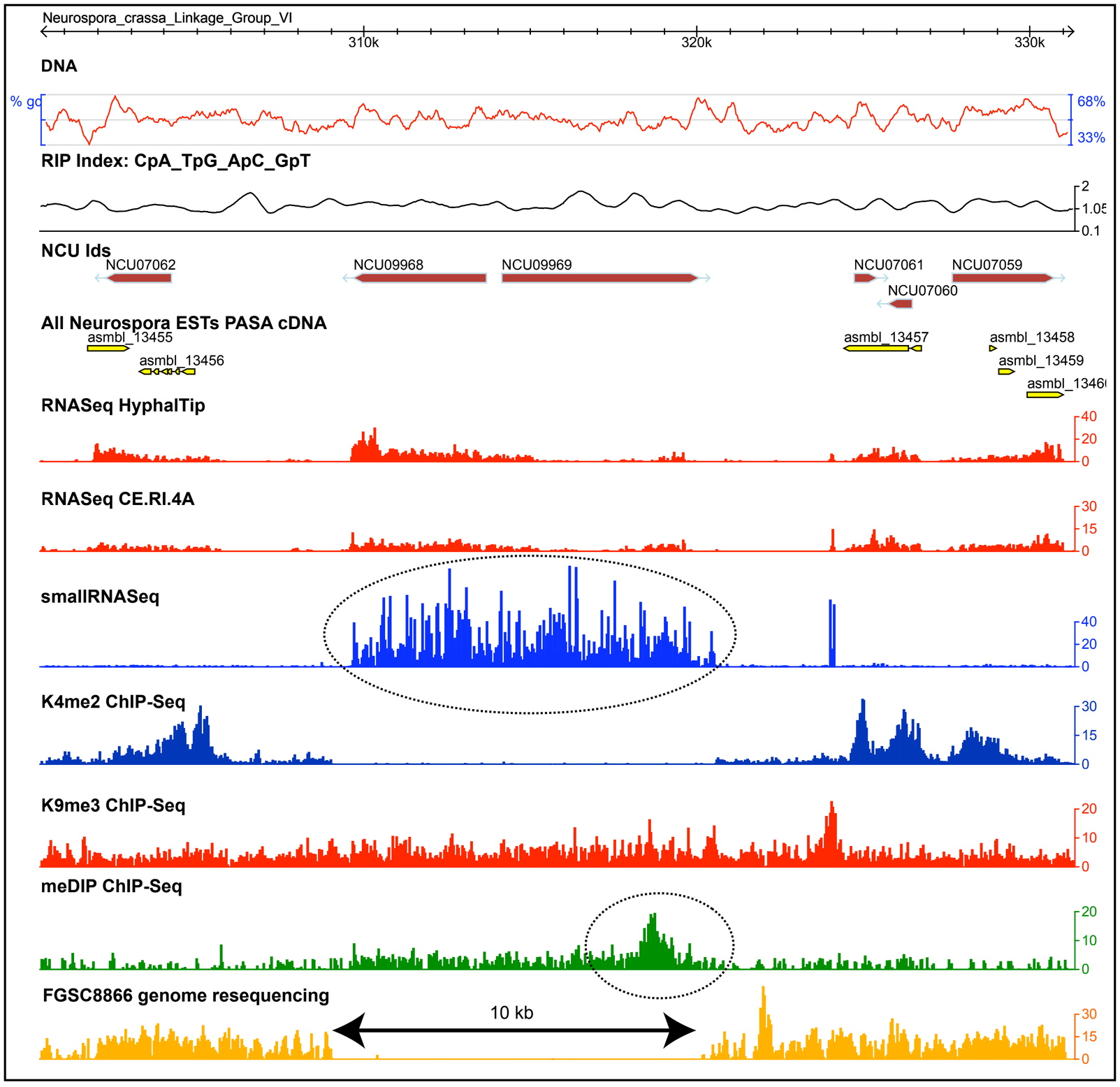
+strand -strand

Intergenic #80



Ethidium staining

Kristina Smith



Small RNA hotspot

Not RIPed

No ESTs

9968 more highly expressed from RNASeq

High smallRNA expression

Region absent in strain from K4 ChIP-Seq

Methylated region

Missing in Tamil Nadu strain

Distribution of hot spot region

- Hot spot missing in other *Neurospora crassa* species.
    - Cannot find copies in NcA or NcC by PCR
- Not found in *N. tetrasperma* or *N. discreta*
- Some similarity found based on translated sequence searches in *Chaetomium* and several Onygenales fungi (*Coccidioides*, *Histoplasma*)
- Hypothesize this is putatively a transposon only found in OR74A.

Conclusions from sequencing and comparative data

- EST and RNA-Seq revealing full transcript diversity
    - Automated updating of UTR regions and alternative splicing forms can detect new genes, splice-sites, and isoforms.
      RNA-Seq more sensitive to transcript level but not-yet refined for all the splice-site detection or alternative splicing
        - A fully-sequence genome is still the beginning in describing possible transcripts and annotation!
- smallRNA-Seq provide additional insight into transcript population
- "smallish" RNA genes, some confirmed by Northerns
- Putative novel transposon in *N. crassa* that is actively silenced not RIPed
- *N. crassa* genome is plastic

Thanks

- John Taylor Lab at UC Berkeley
    - **Chris Ellison**, Thomas Sharpton
- Michael Freitag Lab at Oregon State University
    - **Kristina Smith**

- http://fungalgenomes.org -
  Blog for news, Wiki for protocols, and Data distribution

Code: http://github.com/hyphaltip
GMOD and Gbrowse: http://gmod.org/

Interested in Fungal Evolutionary Genomics?

- My lab will be starting at University of California, Riverside July 2009 http://stajichlab.fungalgenomes.org/

- Interested in genomics, evolution, & fungi?

- Post-transcriptional gene regulation in filamentous fungi

- Fungal cell evolution in early branching chytrid fungi

- Evolution of development in fungi

- Developing bioinformatics and genome informatics tools