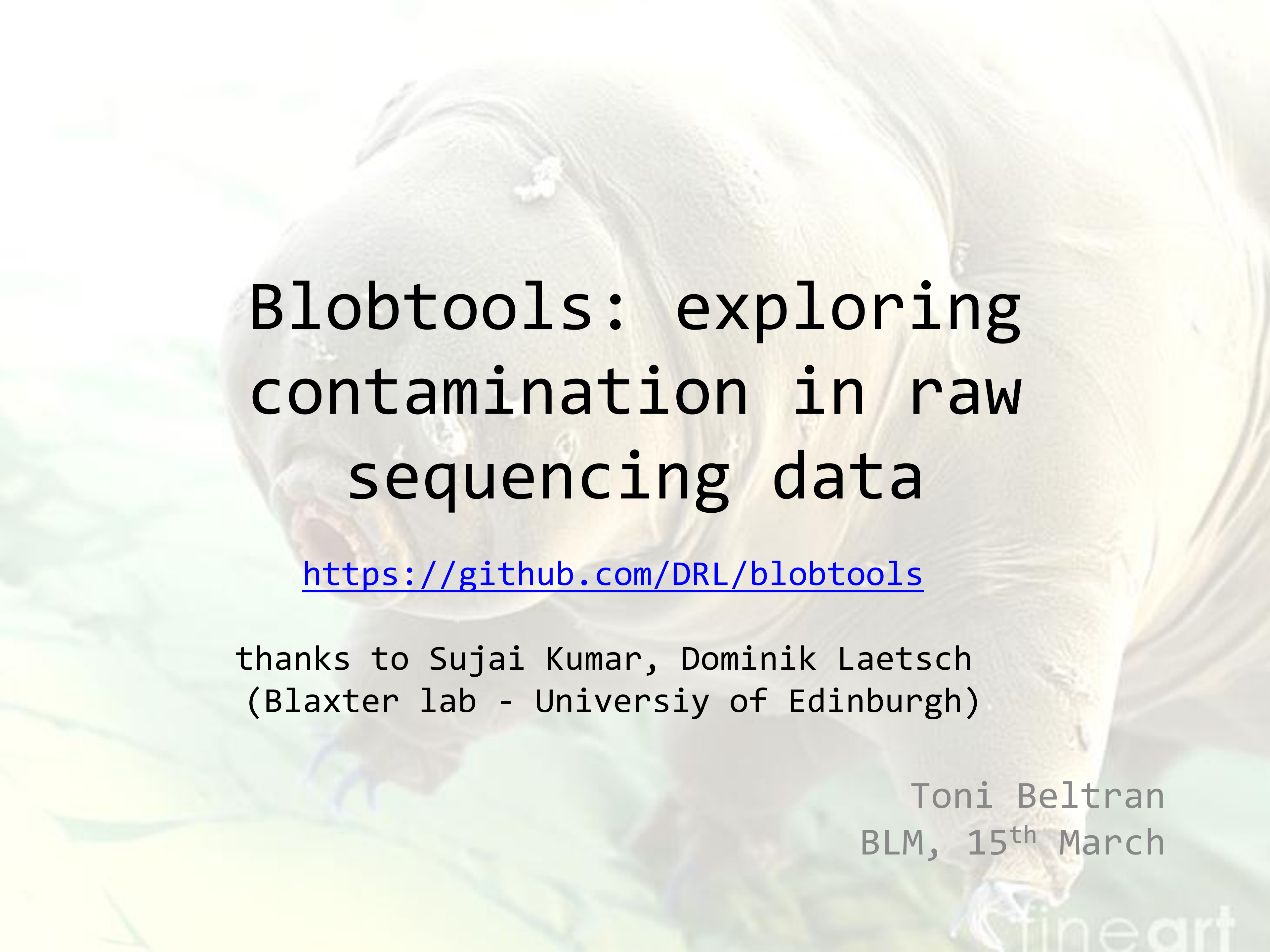


Blobtools: exploring contamination in raw sequencing data

https://github.com/DRL/blobtools

thanks to Sujai Kumar, Dominik Laetsch
(Blaxter lab - Universiy of Edinburgh)

Toni Beltran
BLM, 15th March

Genome assembly is an attempt to accurately represent an entire genome sequence from a large set of very short DNA sequences

Genome assembly is an attempt to accurately represent an entire genome sequence from a large set of very short DNA sequences

“A tremendous amount of genome analysis is built upon the framework of the DNA sequence itself: not only are **genes and regulatory sites** anchored in the sequence, but analyses of **synteny**, **duplications** and **evolutionary relationships** among species all depend on having the correct structure of the genome. We need to devote more effort to making sure the basis for all these analyses does not turn out to be a house of cards.”

Salzberg and Yorke, 2005.

“A tremendous amount of genome analysis is built upon the framework of the DNA sequence itself: not only are **genes and regulatory sites** anchored in the sequence, but analyses of **synteny**, **duplications** and **evolutionary relationships** among species all depend on having the correct structure of the genome. We need to devote more effort to making sure the basis for all these analyses does not turn out to be a house of cards.”

Salzberg and Yorke, 2005.

With the democratisation of sequencing technologies, this is more relevant now than ever.

Genome assembly is a hard problem:

Repeats

Polymorphism

Sequencing errors and biases

Computational requirements

Contamination

Genome assembly is a hard problem:

Repeats

Polymorphism

Sequencing errors and biases

Computational requirements

Contamination

Contamination in sequencing datasets

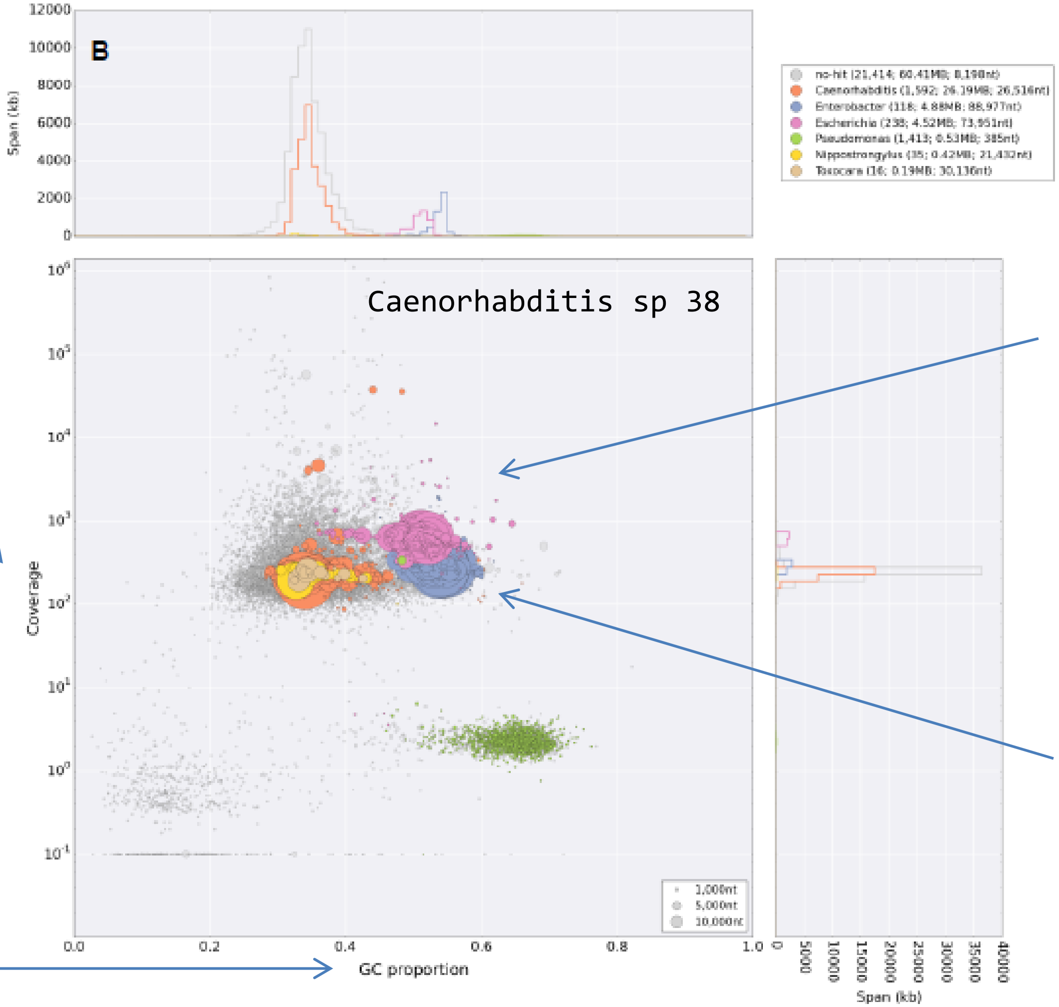
Small target organisms:
need to pool several
individuals

Sequencing data will include
“food” and symbiotic microbiota

Contaminant contigs will
interfere with downstream
analysis

Contaminants can compromise the
assembly of the target genome

What is a “blob plot”?



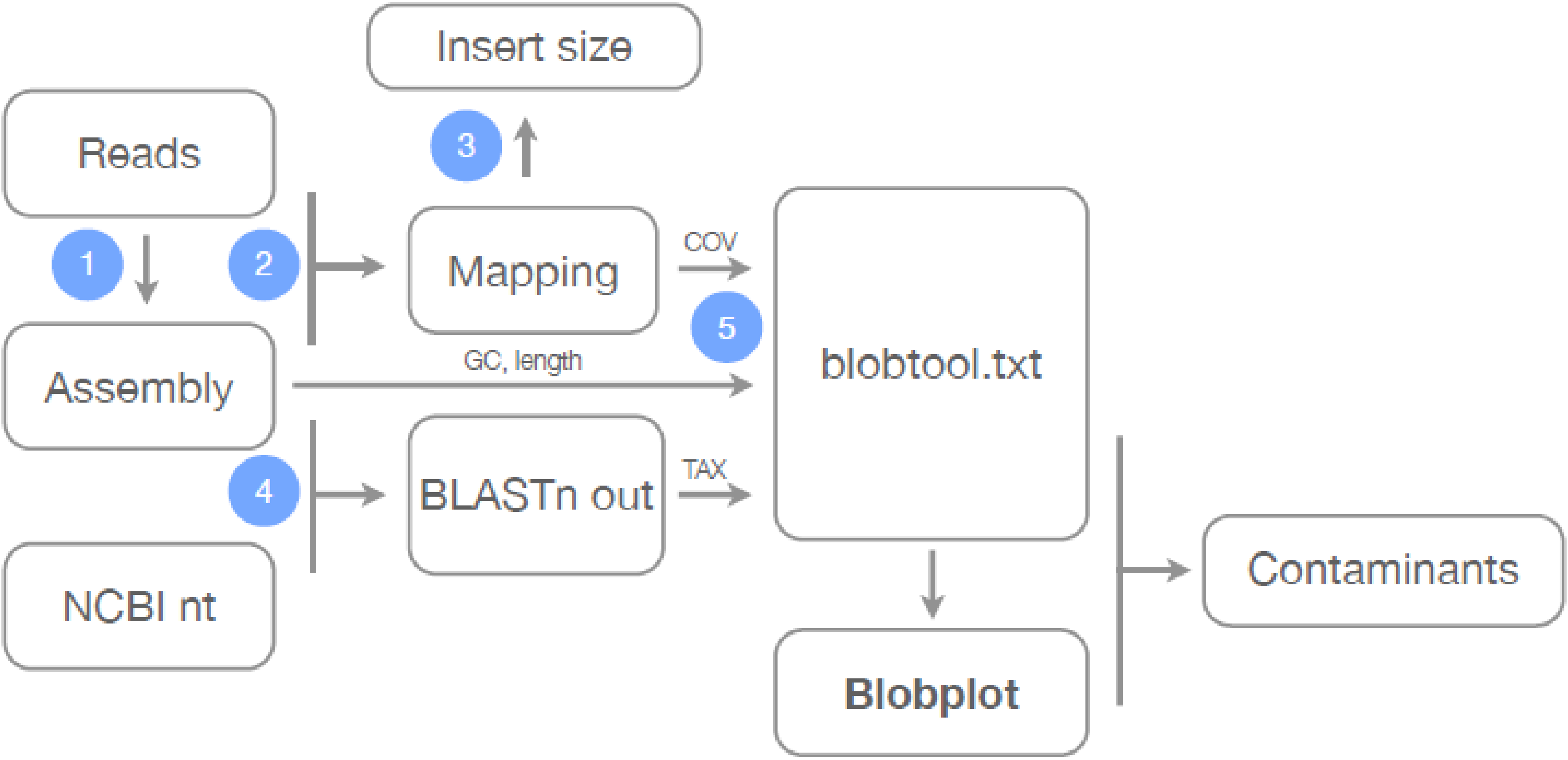
Proxy of **molarity** in the input DNA

Taxonomic annotation in colour

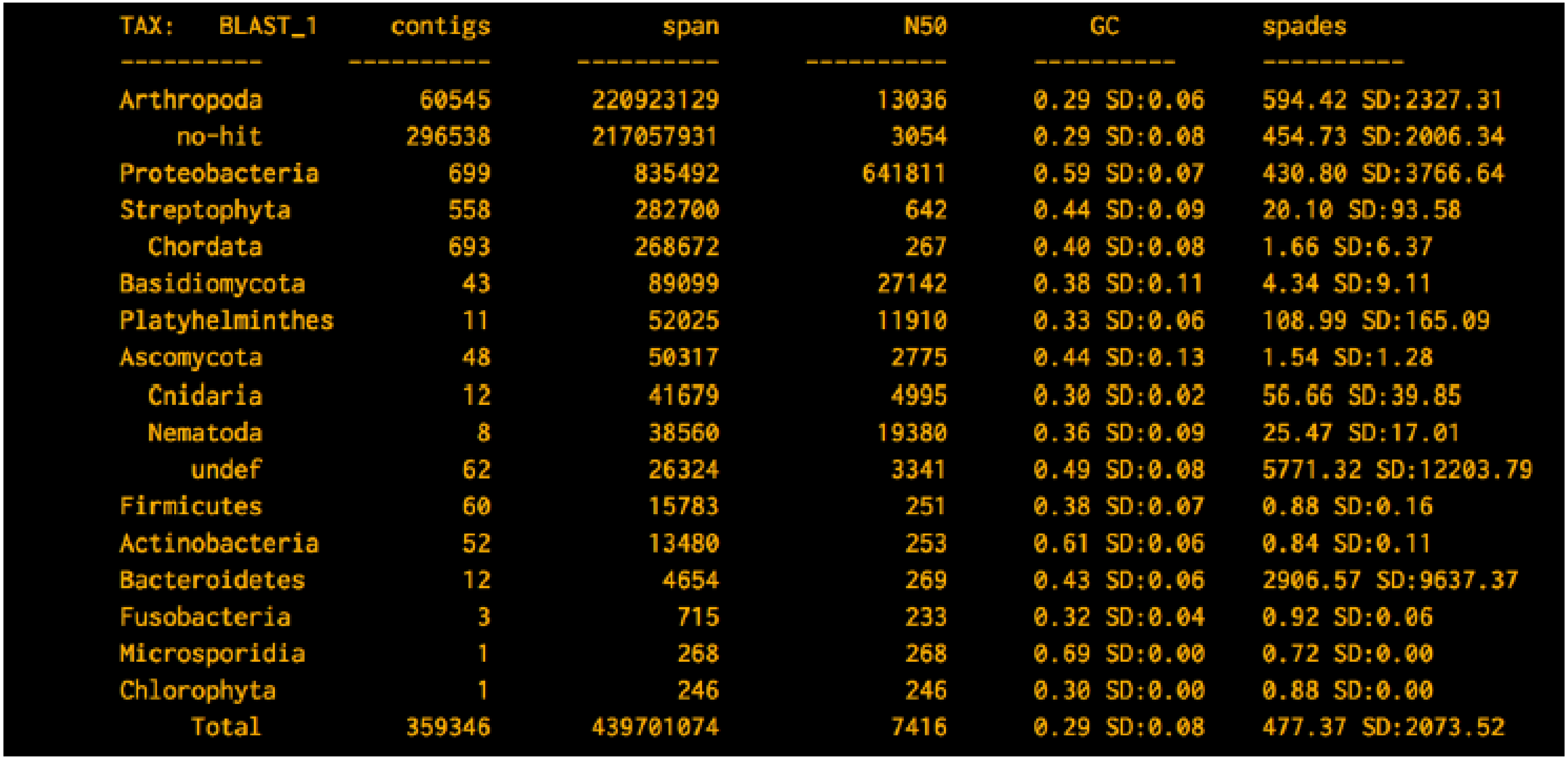
The size of the blob represents the length of the contig

Proxy for **species membership**

How to make a “blob plot”

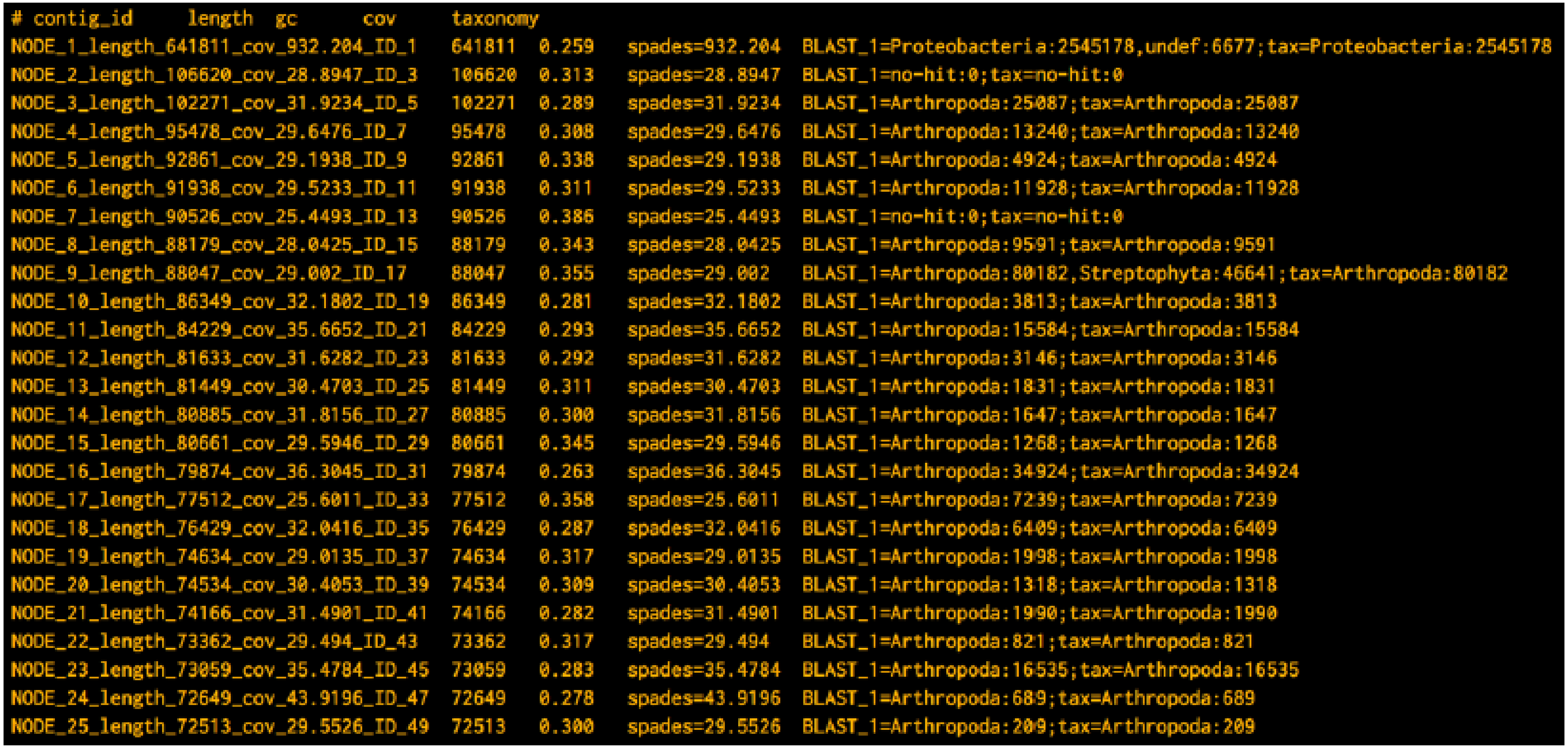


Blobplot.stats.txt



```
TAX:   BLAST_1     contigs            span             N50          GC          spades
----------      ----------      ----------      ----------      ----------      ----------
Arthropoda           60545       220923129           13036      0.29 SD:0.06    594.42 SD:2327.31
    no-hit          296538       217057931            3054      0.29 SD:0.08    454.73 SD:2006.34
Proteobacteria         699          835492          641811      0.59 SD:0.07    430.80 SD:3766.64
Streptophyta           558          282700             642      0.44 SD:0.09    20.10 SD:93.58
  Chordata             693          268672             267      0.40 SD:0.08    1.66 SD:6.37
Basidiomycota           43           89099           27142      0.38 SD:0.11    4.34 SD:9.11
Platyhelminthes         11           52025           11910      0.33 SD:0.06    108.99 SD:165.09
Ascomycota              48           50317            2775      0.44 SD:0.13    1.54 SD:1.28
  Cnidaria              12           41679            4995      0.30 SD:0.02    56.66 SD:39.85
  Nematoda               8           38560           19380      0.36 SD:0.09    25.47 SD:17.01
     undef              62           26324            3341      0.49 SD:0.08    5771.32 SD:12203.79
Firmicutes              60           15783             251      0.38 SD:0.07    0.88 SD:0.16
Actinobacteria          52           13480             253      0.61 SD:0.06    0.84 SD:0.11
Bacteroidetes           12            4654             269      0.43 SD:0.06    2906.57 SD:9637.37
Fusobacteria             3             715             233      0.32 SD:0.04    0.92 SD:0.06
Microsporidia            1             268             268      0.69 SD:0.00    0.72 SD:0.00
Chlorophyta              1             246             246      0.30 SD:0.00    0.88 SD:0.00
     Total          359346       439701074            7416      0.29 SD:0.08    477.37 SD:2073.52
```

Blobplot.txt



```
# contig_id     length  gc      cov     taxonomy
NODE_1_length_641811_cov_932.204_ID_1   641811  0.259   spades=932.204  BLAST_1=Proteobacteria:2545178,undef:6677;tax=Proteobacteria:2545178
NODE_2_length_106620_cov_28.8947_ID_3   106620  0.313   spades=28.8947  BLAST_1=no-hit:0;tax=no-hit:0
NODE_3_length_102271_cov_31.9234_ID_5   102271  0.289   spades=31.9234  BLAST_1=Arthropoda:25087;tax=Arthropoda:25087
NODE_4_length_95478_cov_29.6476_ID_7    95478   0.308   spades=29.6476  BLAST_1=Arthropoda:13240;tax=Arthropoda:13240
NODE_5_length_92861_cov_29.1938_ID_9    92861   0.338   spades=29.1938  BLAST_1=Arthropoda:4924;tax=Arthropoda:4924
NODE_6_length_91938_cov_29.5233_ID_11   91938   0.311   spades=29.5233  BLAST_1=Arthropoda:11928;tax=Arthropoda:11928
NODE_7_length_90526_cov_25.4493_ID_13   90526   0.386   spades=25.4493  BLAST_1=no-hit:0;tax=no-hit:0
NODE_8_length_88179_cov_28.0425_ID_15   88179   0.343   spades=28.0425  BLAST_1=Arthropoda:9591;tax=Arthropoda:9591
NODE_9_length_88047_cov_29.002_ID_17    88047   0.355   spades=29.002   BLAST_1=Arthropoda:80182,Streptophyta:46641;tax=Arthropoda:80182
NODE_10_length_86349_cov_32.1802_ID_19  86349   0.281   spades=32.1802  BLAST_1=Arthropoda:3813;tax=Arthropoda:3813
NODE_11_length_84229_cov_35.6652_ID_21  84229   0.293   spades=35.6652  BLAST_1=Arthropoda:15584;tax=Arthropoda:15584
NODE_12_length_81633_cov_31.6282_ID_23  81633   0.292   spades=31.6282  BLAST_1=Arthropoda:3146;tax=Arthropoda:3146
NODE_13_length_81449_cov_30.4703_ID_25  81449   0.311   spades=30.4703  BLAST_1=Arthropoda:1831;tax=Arthropoda:1831
NODE_14_length_80885_cov_31.8156_ID_27  80885   0.300   spades=31.8156  BLAST_1=Arthropoda:1647;tax=Arthropoda:1647
NODE_15_length_80661_cov_29.5946_ID_29  80661   0.345   spades=29.5946  BLAST_1=Arthropoda:1268;tax=Arthropoda:1268
NODE_16_length_79874_cov_36.3045_ID_31  79874   0.263   spades=36.3045  BLAST_1=Arthropoda:34924;tax=Arthropoda:34924
NODE_17_length_77512_cov_25.6011_ID_33  77512   0.358   spades=25.6011  BLAST_1=Arthropoda:7239;tax=Arthropoda:7239
NODE_18_length_76429_cov_32.0416_ID_35  76429   0.287   spades=32.0416  BLAST_1=Arthropoda:6409;tax=Arthropoda:6409
NODE_19_length_74634_cov_29.0135_ID_37  74634   0.317   spades=29.0135  BLAST_1=Arthropoda:1998;tax=Arthropoda:1998
NODE_20_length_74534_cov_30.4053_ID_39  74534   0.309   spades=30.4053  BLAST_1=Arthropoda:1318;tax=Arthropoda:1318
NODE_21_length_74166_cov_31.4901_ID_41  74166   0.282   spades=31.4901  BLAST_1=Arthropoda:1990;tax=Arthropoda:1990
NODE_22_length_73362_cov_29.494_ID_43   73362   0.317   spades=29.494   BLAST_1=Arthropoda:821;tax=Arthropoda:821
NODE_23_length_73059_cov_35.4784_ID_45  73059   0.283   spades=35.4784  BLAST_1=Arthropoda:16535;tax=Arthropoda:16535
NODE_24_length_72649_cov_43.9196_ID_47  72649   0.278   spades=43.9196  BLAST_1=Arthropoda:689;tax=Arthropoda:689
NODE_25_length_72513_cov_29.5526_ID_49  72513   0.300   spades=29.5526  BLAST_1=Arthropoda:209;tax=Arthropoda:209
```

Remove contaminant reads

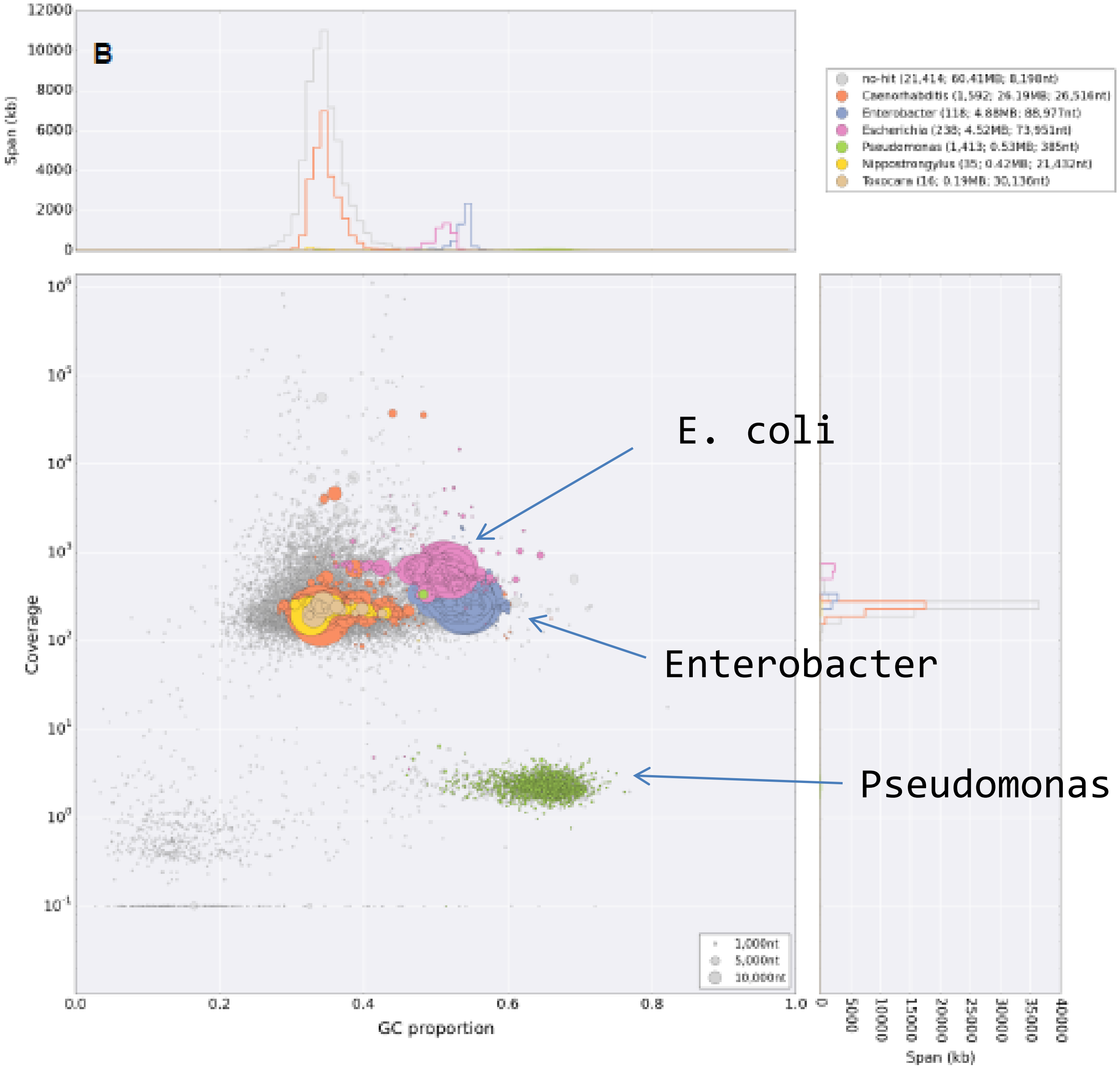
If we can identify the contaminants directly, and they have been sequenced, remove reads mapping to their genomes.

If not, filter contigs based on GC content, coverage and taxonomic information.
-Remove reads mapping to those contigs
-Reassemble until no contaminant contigs are found

Remove contaminant reads

If we can identify the contaminants directly, and they have been sequenced, remove reads mapping to their genomes.

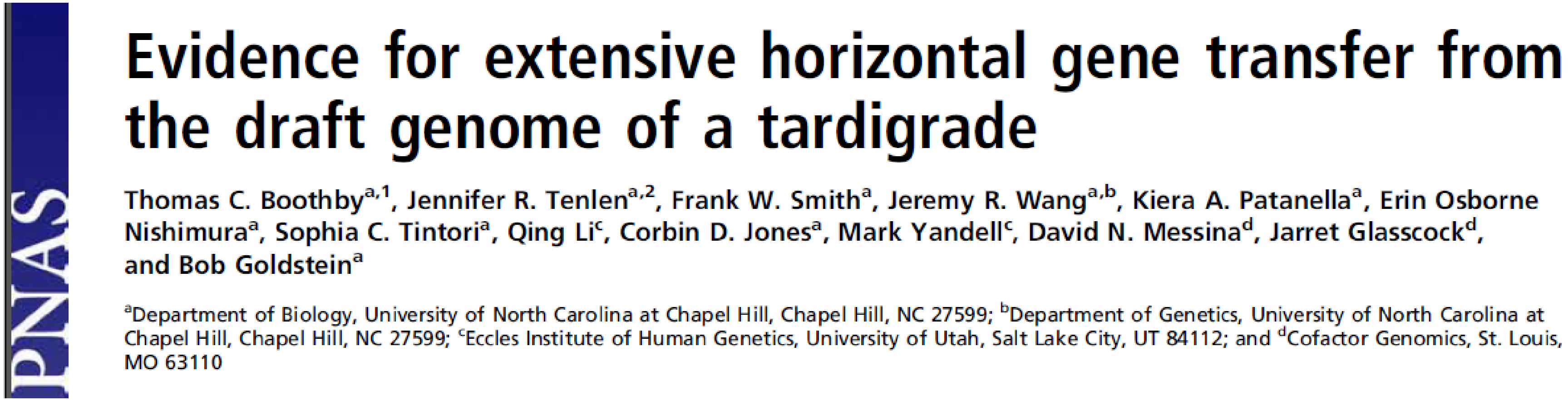
If not, filter contigs based on GC content, coverage and taxonomic information.
-Remove reads mapping to those contigs
-Reassemble until no contaminant contigs are found



E. coli

Enterobacter

Pseudomonas

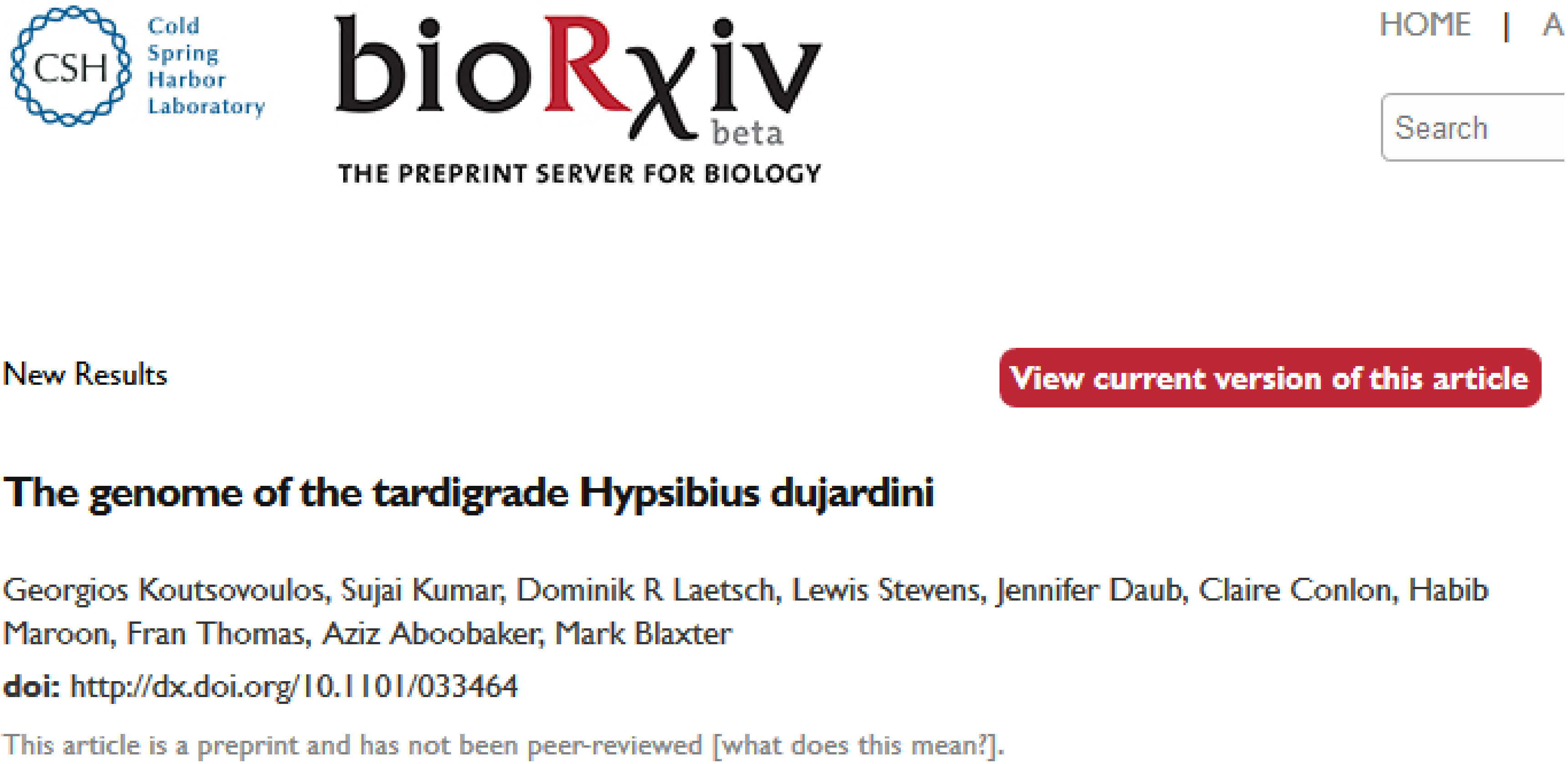


Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade

Thomas C. Boothby[a,1], Jennifer R. Tenlen[a,2], Frank W. Smith[a], Jeremy R. Wang[a,b], Kiera A. Patanella[a], Erin Osborne Nishimura[a], Sophia C. Tintori[a], Qing Li[c], Corbin D. Jones[a], Mark Yandell[c], David N. Messina[d], Jarret Glasscock[d], and Bob Goldstein[a]

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [b]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [c]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112; and [d]Cofactor Genomics, St. Louis, MO 63110

“Genome sequencing, direct confirmation of physical linkage, and phylogenetic analysis revealed that a large fraction of the *H. dujardini* genome is derived from diverse bacteria as well as plants, fungi, and Archaea. We estimate that **approximately one-sixth of tardigrade genes entered by HGT**, nearly double the fraction found in the most extreme cases of HGT into animals known to date.”



CSH Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY
HOME | A
Search
View current version of this article

New Results

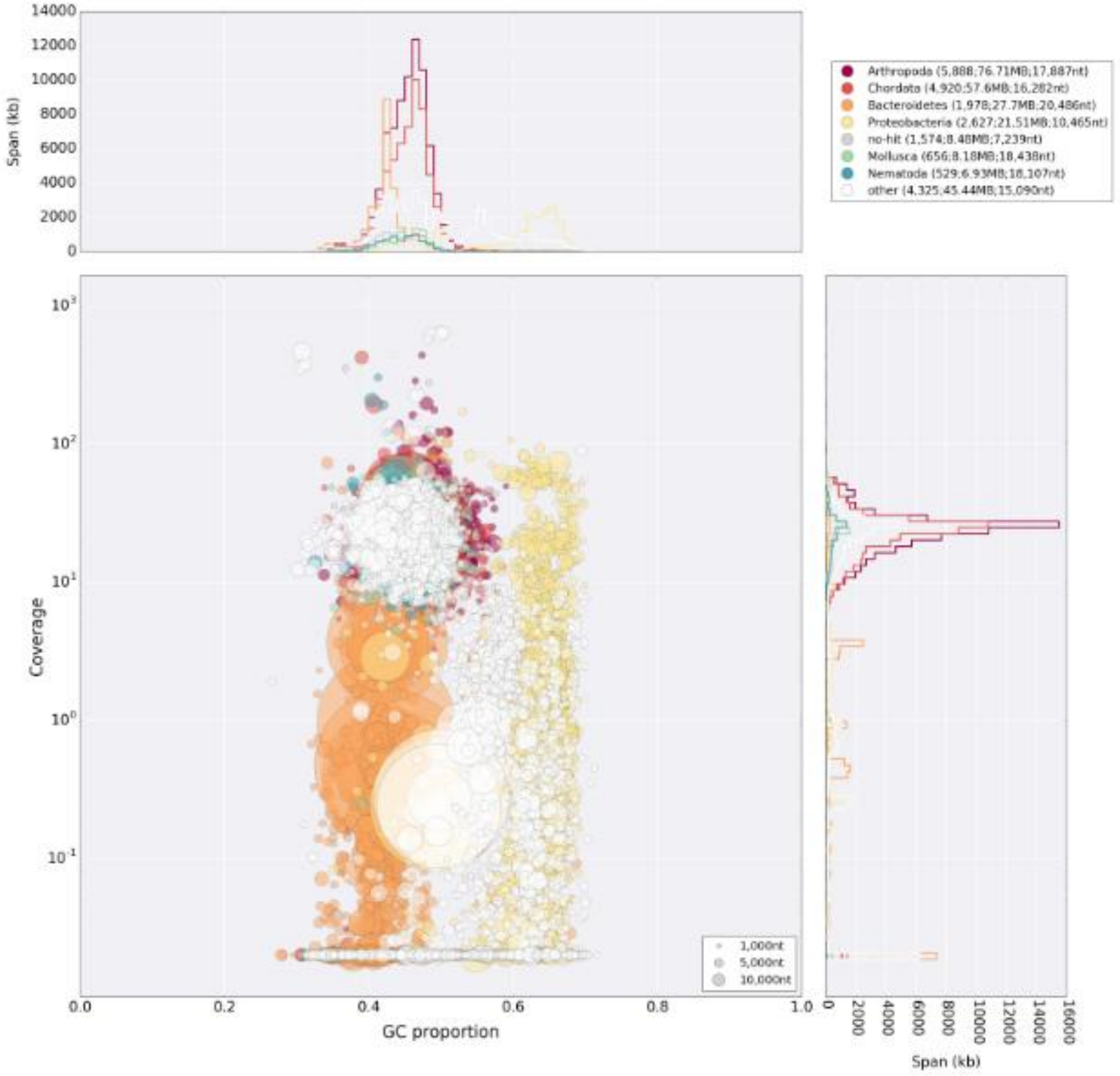
The genome of the tardigrade Hypsibius dujardini

Georgios Koutsovoulos, Sujai Kumar, Dominik R Laetsch, Lewis Stevens, Jennifer Daub, Claire Conlon, Habib Maroon, Fran Thomas, Aziz Aboobaker, Mark Blaxter

doi: http://dx.doi.org/10.1101/033464

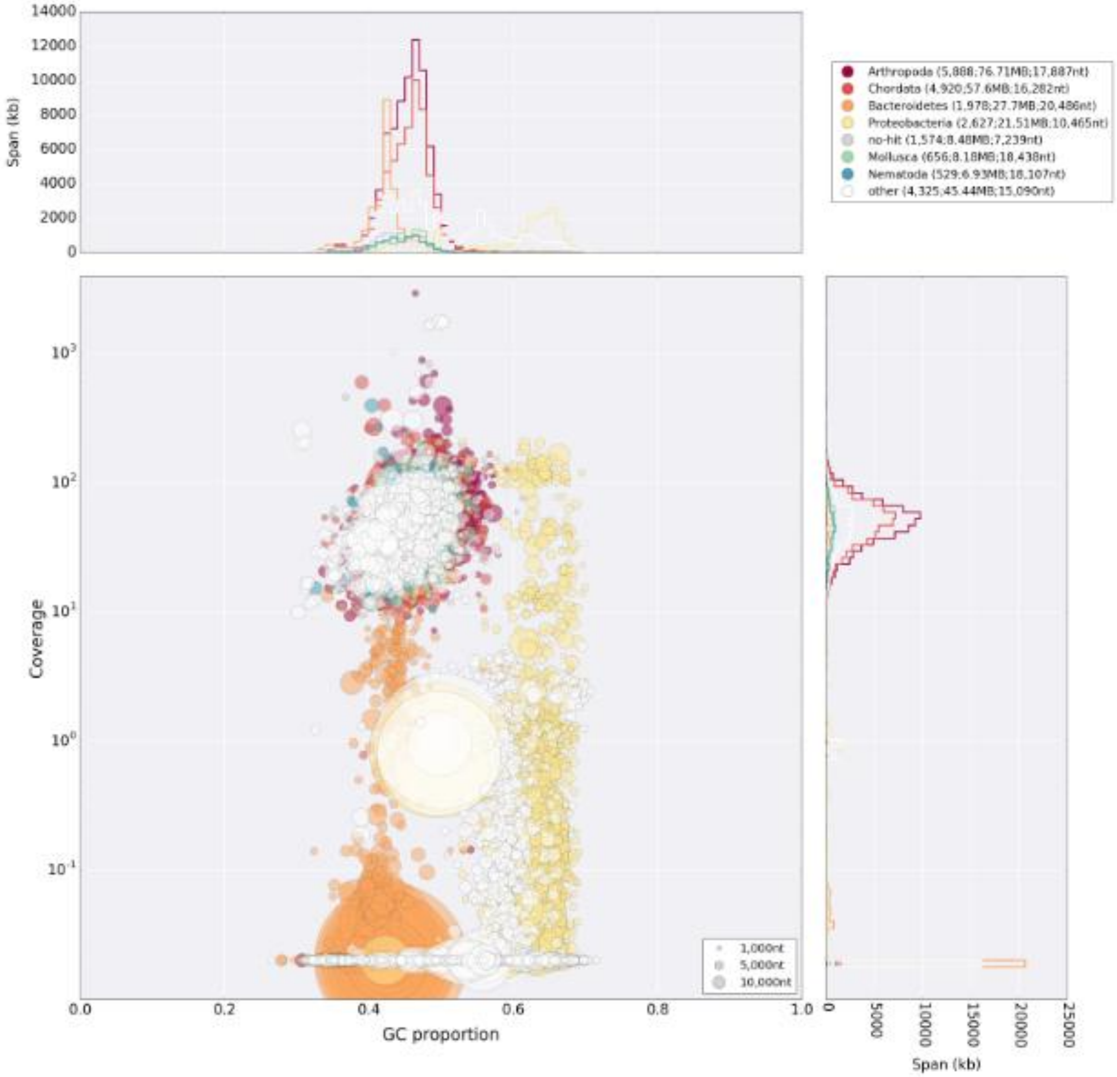
This article is a preprint and has not been peer-reviewed [what does this mean?].

UNC raw sequencing data shows lots of contigs with low/no coverage



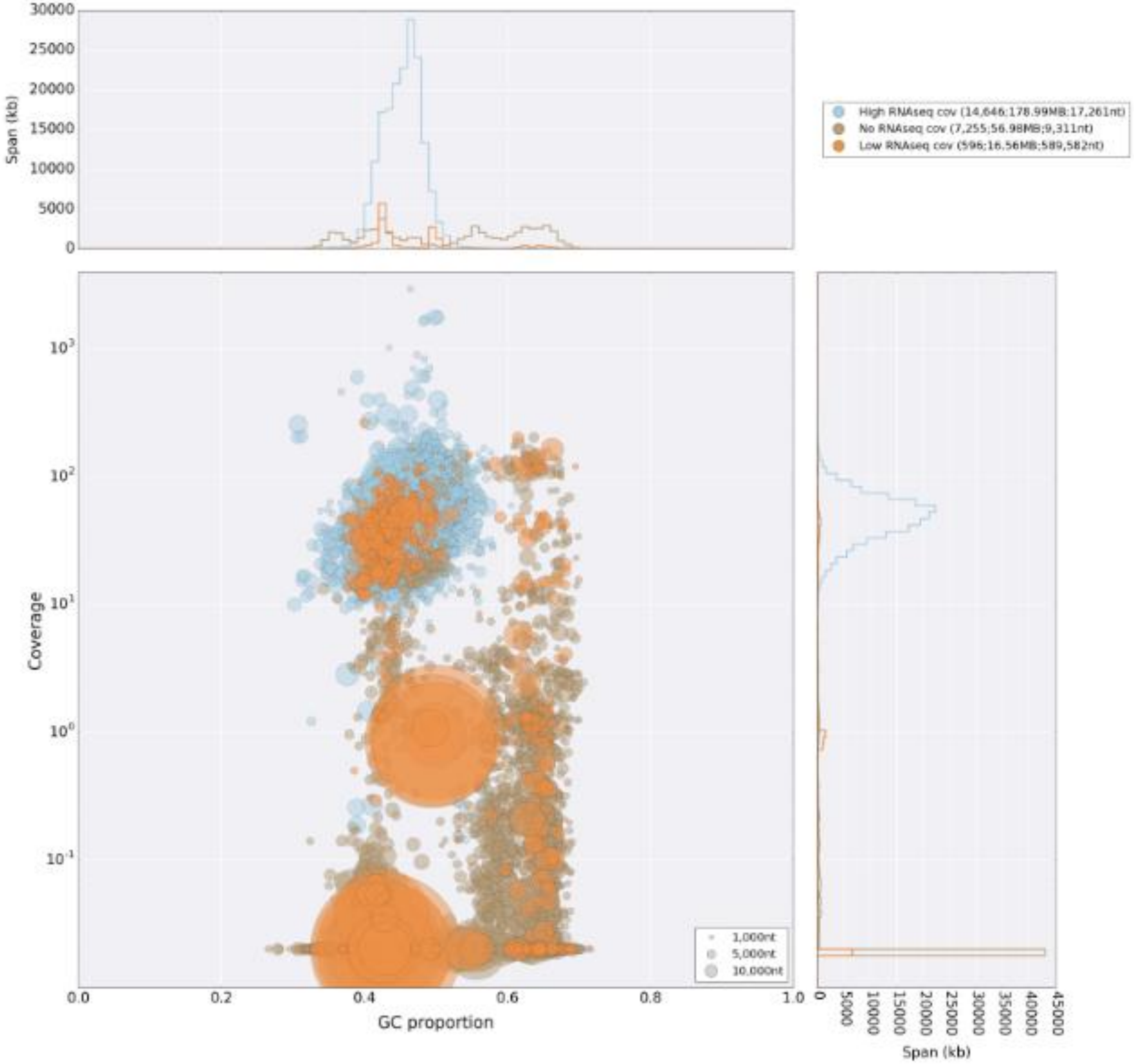
Koutsovoulos et. al. 2016

Edinburgh independent sequencing shows lots of contigs with low/no coverage



Koutsovoulos et. al. 2016

Contigs with low coverage are not represented in independent RNA-seq data



Koutsovoulos et. al. 2016

You should regard every draft genome assembly as work in progress.

In some years time we will look back at genome assembly at this time with embarrassment – but this is the best we can do now.

We should be more strict evaluating genome assembly quality. Check contamination even in published genome assemblies!

There are reasons to be optimistic (long read technologies, single chromosome sequencing, Hi-C).

Open science is fast and effective.