Spring 2013 Genomics Exam #1
Genomic Sequences

There is no time limit on this test, though I don't want you to spend too much time on it. I have tried to design an exam that will take less time that exams in the past. You do not need to read any additional papers other than the ones I send to you. There are 4 pages, including this cover sheet, for this test. There are no Discovery Questions on this exam. You are not allowed discuss the test with anyone until all exams are turned in at 10:30 am on Wednesday February 13. **ELECTRONIC COPIES OF YOUR EXAM ANSWERS ARE DUE AT 10:30 am ON WEDNESDAY FEBRUARY 13**. You may use a calculator, a ruler, your notes, the book, and the internet. You may work on this exam in as many blocks of time as you want. Submit your electronic version before 10:30 am (eastern time zone).

The **answers to the questions must be typed in a Word file and emailed to me as an attachment**. Be sure to backup your test answers just in case (I suggest a thumb drive or other removable medium). You will need to capture screen images as a part of your answers which you may do without seeking permission since your test answers will not be in the public domain. Remember to explain your thoughts in *your* own words and use screen shots to support your answers. Screen shots without ***your*** words are worth very few points. Support your answers with data using screen shots liberally.

DO NOT READ or DOWNLOAD ANY NEW PAPERS FOR THIS EXAM. You may search and read abstracts. RELY ON YOUR EXPERIENCE, AND YOUR SKILLS. Spell out your logic for each answer.

-3 pts if you do not follow this direction.
Please do not write or type your name on any page other than this cover page.
Staple all your pages (INCLUDING THE TEST PAGES) together when finished with the exam.

Name (please type):

Write out the full pledge and sign (electronic signature is ideal):

How long did this exam take you to complete?

20 pts

1) I want you to use a database you have never seen before called EuPathDB (http://eupathdb.org/eupathdb/). Use the Giardia portion of this integrated database. We will not use the full power of what this site can do, but you will get a sense of its potential as you work on this problem.

Your task is to identify a protein target for a drug to be developed by a company called Mayking Itup, LLC. You will have to figure out how to use EuPathDB to answer most of these questions.

a) What is Giardia and what sort of disease does it cause? Support your answer by providing the URL of your information source(s). **Limit your answer to 3 sentences or less.**

b) Identify a set of proteins whose features include a known epitope and it is an integral membrane protein. You also must be certain that the protein is expressed by using EST data. How many proteins are in this set of genes/proteins? Provide a screen shot to support your answer.

c) Choose one gene/protein from your list above that you would like to inhibit based on its biological process. Name that gene by its common name and its DB accession number.

d) How many transmembrane domains are predicted for your chosen protein? How many amino acids in this protein? Support your answer with data.

e) Find another way to independently confirm via computer (prediction) whether the number of transmembrane domains you found for part (d) is correct or not. Support your answer with data.

20 pts

2) This time, I want you to use the JCVI CMR (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi). Your task is to compare the predicted metabolic pathways for converting acetate into CO_2, NADH, $FADH_2$ and ATP in two strains of *E. coli*: 1) the first non-pathogenic strain to have its genome sequenced and 2) the pathogenic strain EDL933.

a) What was the original source of strain EDL933? Tell me where you found your answer.

b) Find a fundamental KEGG biochemical pathway in CMR that shows a difference in metabolic capacity between these two strains. You should be looking for a pathway where each genome has at least one enzyme the other lacks. Support your answer with a screenshot showing the differences.

c) Do you accept that the key metabolic pathway for both strains is accurately annotated in this database? Explain your reasoning. **Limit your answer to 3 sentences or less.**

d) Choose one enzyme that is found in only one strain in your screen shot from JCVI CMR and determine if CMR is correct or not about it being absent from the other strain. You will have to tell me where you searched and how you conducted the search. If you can disprove the map, support your answer with data. If you cannot disprove the CMR map, explain why you were unable to find the answer.

20 pts

3) Horizontal gene transfer is sometimes called lateral gene transfer (LGT) in order to keep the typical undergraduate confused and to provide yet another TLA. However, you are now on the inside crowd, so I want to ask you some questions about LGT.

a) Below is a figure from a paper that claims to have evidence of bacteria to eukaryote LGT. The method combined fluorescent labeling of a chromosome with FISH. Evaluate these data and tell me if

you think the evidence is either 1) inconsistent with LGT, 2) consistent with LGT, 3) compelling evidence of LGT or 4) inconclusive. Support your answer with data. **Limit your answer to 3 sentences maximum.**

consistent with, but controls are missing

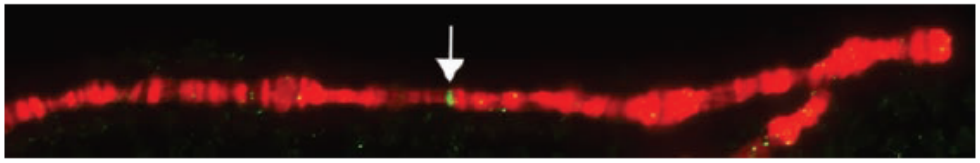


Fig. 1. Fluorescence microscopy evidence supporting *Wolbachia*/host LGT. DNA in the polytene chromosomes of *D. ananassae* were stained with propidium iodide (red), whereas a probe for the *Wolbachia* gene WD_0484 bound to a unique location (green, arrow) on chromosome 2L.

b) The accompanying PDF file called Figure2.pdf contains some data based on genome sequence analysis. The authors present Figure 2 as supporting evidence that LGT has occurred. Your task is to evaluate the data and tell me if you think the data are 1) inconsistent with LGT, 2) consistent with LGT, 3) compelling evidence of LGT or 4) inconclusive. Support your assessment by citing data appropriately. Assume zero sequencing or assembly errors happened in this research. **Limit your answer to 5 sentences maximum.**

c) The accompanying PDF file called Figure3.pdf contains additional data. The table shows 9 examples of LGT. Rank each of the nine examples from most compelling to least compelling and explain your reason for each ranking in one sentence maximum. If you feel a tie is required, then list multiple examples for a single number and reduce the final number accordingly.

20 pts

4) Here is a sequence of DNA. You need to answer the following questions using this sequence as your starting place. `agtttttcacatatctccatcgcctcagttgctatcaaca`

a) From which gene and species did this sequence come? Support your answer with evidence and be as accurate as you can be with your answer.

b) How many exons are in the human ortholog? What is this gene's chromosomal position in humans? Support your answer with data.

c) List (numbered list) as many biological processes as you can find for the human ortholog. Are all of the processes very similar, or do you see some pretty diverse categories? **Explain your answer in 1 sentence.**

d) Propose a reasonable model to show how mutant alleles of this gene could be passed down from fathers but <u>not</u> mothers. This answer is not intended to be a generic one that could apply to any gene, but you should combine what we have learned in class with the specific role(s) of this protein.

e) Demonstrate the degree of sequence conservation between the human ortholog and other species. You must provide a screenshot and then write a summary of what you conclude from your screen shot. **Limit your summary to a maximum of 2 sentences.**

20 pts

5) This last question has to do with the ENCODE project. *You are NOT allowed to look up any scientific ENCODE papers or ENCODE abstracts*. Therefore, do not perform a PubMed search. If you do a Google search, be sure to screen any hits before clicking to be sure you are NOT reading a scientific paper or abstract.

a) What was the purpose of the ENCODE project? **Limit your answer to a maximum of 2 sentences.**

b) How many ENCODE papers were published in a coordinated way in late 2012? **Limit your answer to a maximum of 1 sentence.**

c) What does DHS stand for in the ENCODE project? Describe the physical characteristic of DNA DHS is measuring. **Limit your answer to a maximum of 1 sentence.**

d) In one sentence, define what TSS means. Support your definition using data from Figure_4.pdf.

e) Summarize the two main lessons in panel B of Figure_4.pdf. **Limit your answer to a maximum of 3 sentences.**

f) This is the first time I have ever seen "violin plots". Summarize panel C in Figure_4.pdf. **Limit your answer to a maximum of 2 sentences.**

g) In panel A of Figure_5.pdf, they tested 19 different cell lines for DHS. Summarize the findings of panels A – C **in six sentences or less (two per panel).**

Panel A:

Panel B:

Panel C:

h) View bases 201,574,325 to 201,591,603 on chromosome 1. Show me a screen shot of this region with DHS data included in your display as well as the degree of conserved bases in 5 diverse vertebrates. Summarize what you see in your screen shot based on what you learned in Figure_4.pdf and Figure_5.pdf. **Limit your summary to a maximum of 2 sentences.**

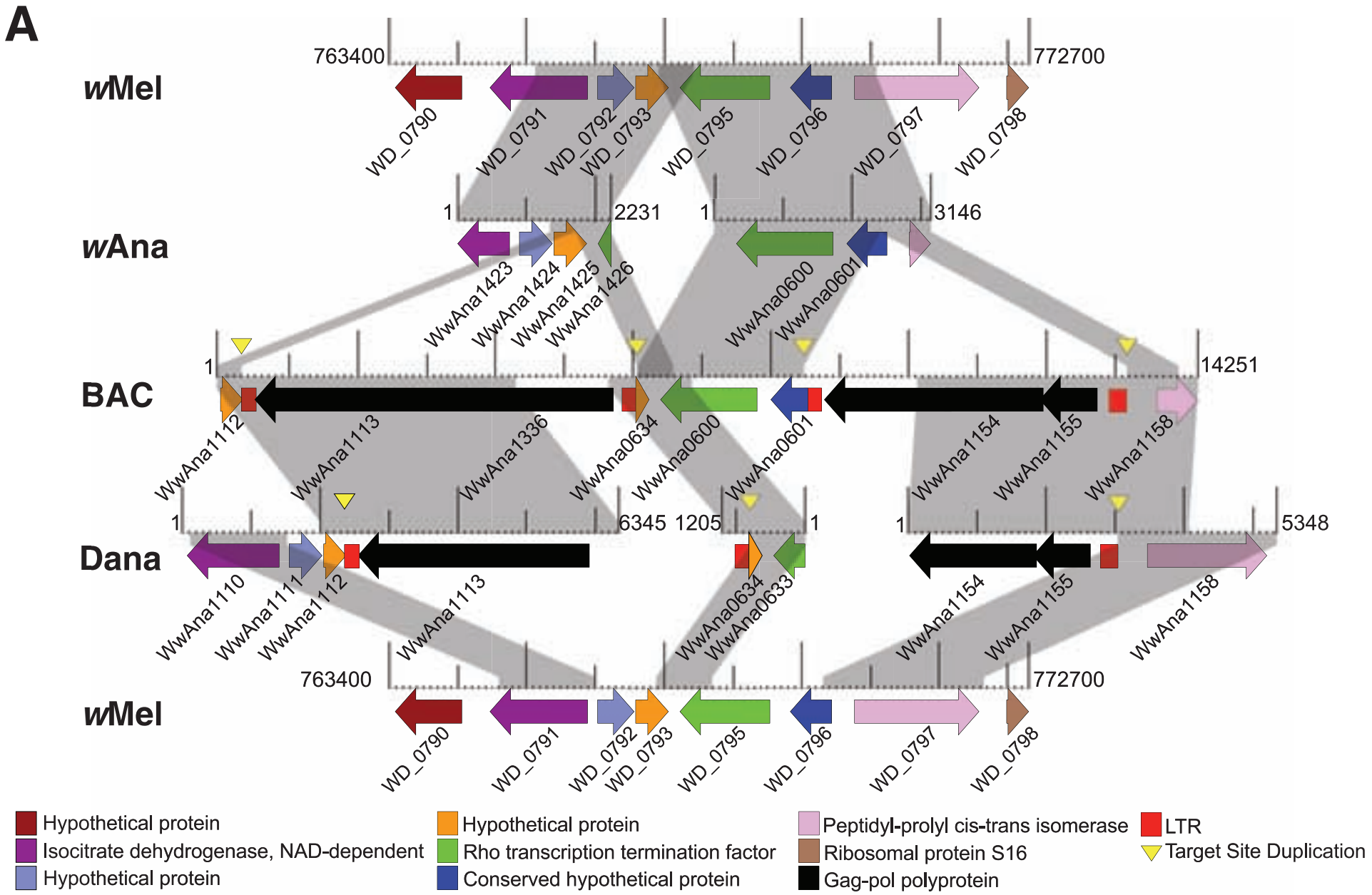
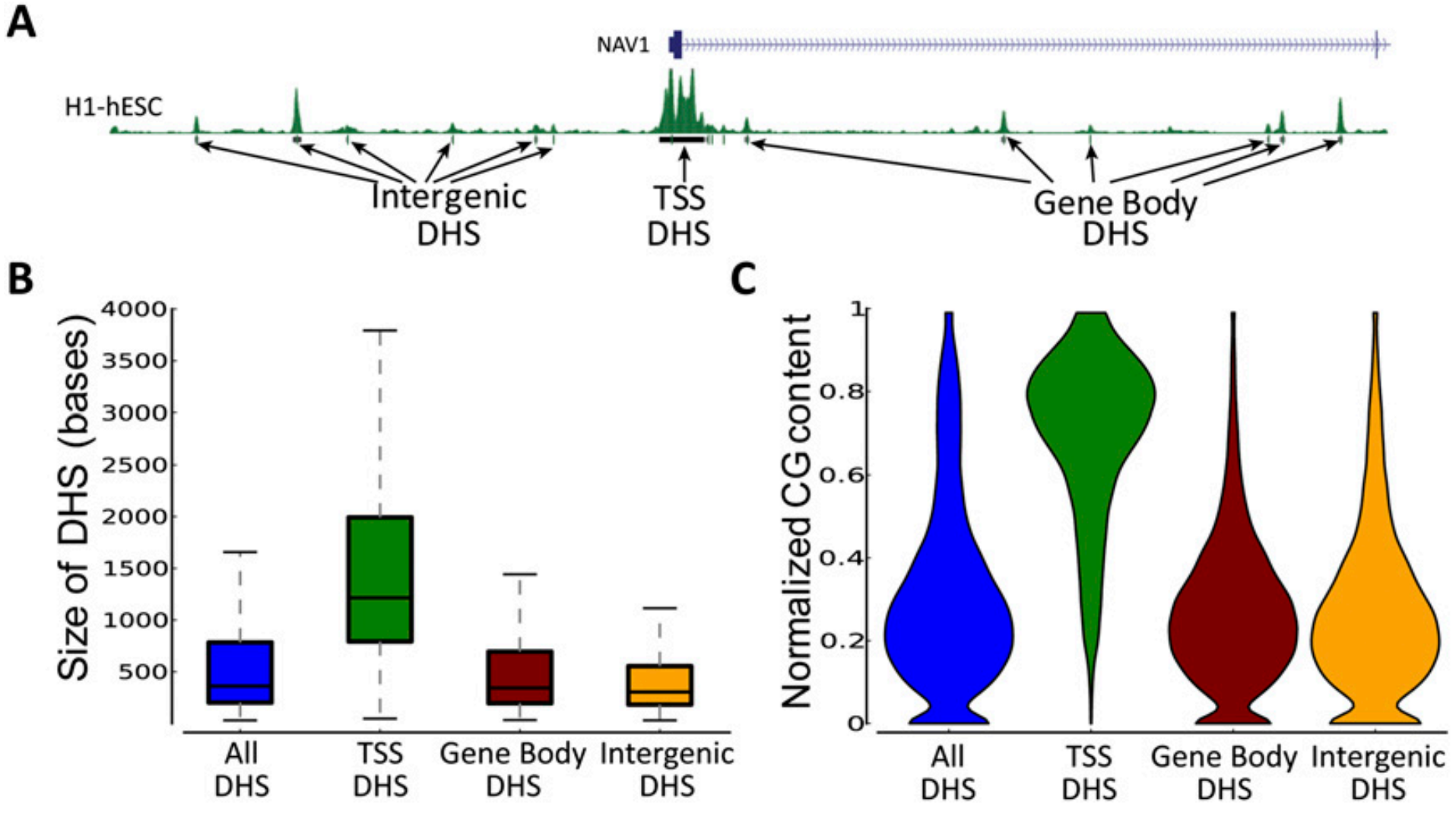
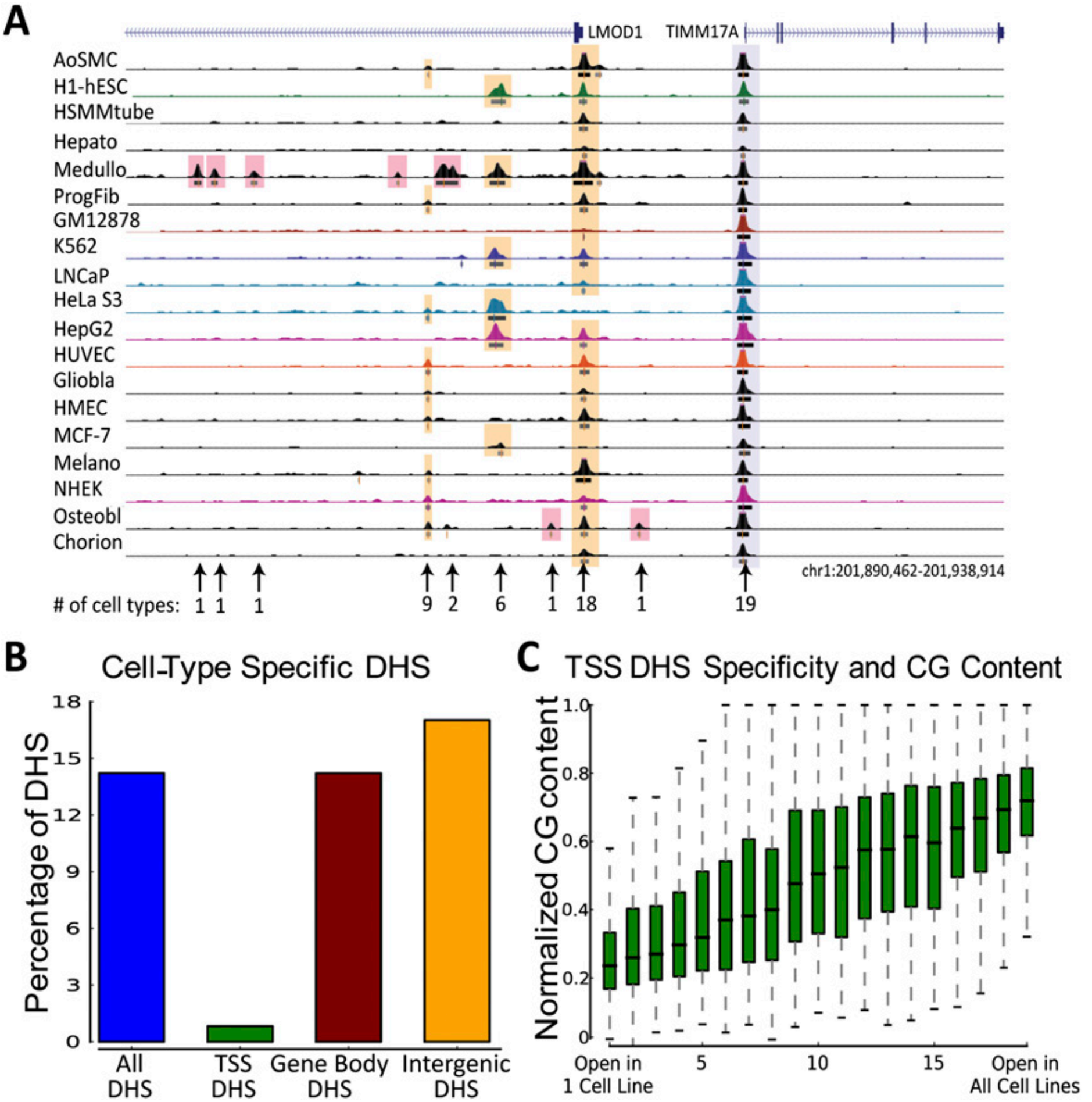

A
wMel
763400
772700
WD_0790
WD_0791
WD_0792
WD_0793
WD_0795
WD_0796
WD_0797
WD_0798
wAna
1
2231
1
3146
WwAna1423
WwAna1424
WwAna1425
WwAna1426
WwAna0600
WwAna0601
BAC
1
14251
WwAna1112
WwAna1113
WwAna1336
WwAna0634
WwAna0600
WwAna0601
WwAna1154
WwAna1155
WwAna1158
Dana
1
6345
1205
1
1
5348
WwAna1110
WwAna1111
WwAna1112
WwAna1113
WwAna0634
WwAna0633
WwAna1154
WwAna1155
WwAna1158
763400
772700
wMel
WD_0790
WD_0791
WD_0792
WD_0793
WD_0795
WD_0796
WD_0797
WD_0798
Hypothetical protein
Isocitrate dehydrogenase, NAD-dependent
Hypothetical protein
Hypothetical protein
Rho transcription termination factor
Conserved hypothetical protein
Peptidyl-prolyl cis-trans isomerase
Ribosomal protein S16
Gag-pol polyprotein
LTR
Target Site Duplication

Table 1. Summary of *Wolbachia* sequences and evidence for LGT in public databases. Junctions were validated by PCR amplification and sequencing (*11*), with the number of successful reactions compared to the number attempted. Species marked with a plus sign are described in the literature as being infected with *Wolbachia*. All whole-genome shotgun sequencing reads were downloaded for 26 arthropod and nematode genomes (*11*). Organisms identified as lacking *Wolbachia* sequences either had no match or matches only to the prokaryotic ribosomal RNA. Because the *Nasonia* genomes are from antibiotic-cured insects, they were identified as having a putative LGT event merely on identification of *Wolbachia* sequences in a read. All other organisms were considered to have putative LGT events if the trace repository contained ≥1 read with (i) >80% nucleotide identity over 10% of the read to a characterized eukaryotic gene, (ii) >80% identity over 10% of the read to a *Wolbachia* gene, and (iii) manual review of the BLAST results for 1 to 20 reads to ensure significance (*11*). NA, not applicable.

Organism	Total traces screened	*Wolbachia* traces	LGT	Junctions validated	*Wolbachia* infection
Trace repository sequences					
Acyrthosiphon pisum (aphid)	4,285,120	0			+
Aedes aegypti (mosquito)	16,238,263	0			–
Anopheles gambiae (mosquito)	5,456,630	0			–
Apis mellifera (honeybee)	3,941,137	0			–
Brugia malayi (filarial nematode)	1,260,214	22,524	+	10/12	+
Culex pipiens quinquefasciatus (mosquito)	7,380,430	21,304	+	0/0	+
Daphnia pulex (crustacean)	2,724,768	0			–
D. ananassae (fruit fly)	3,878,537	38,605	+	6/7	+
D. erecta (fruit fly)	2,916,936	0			–
D. grimshawi (fruit fly)	2,874,111	0			–
D. melanogaster (fruit fly)	1,001,855	0			+
D. mojavensis (fruit fly)	3,130,180	107*	–		–
D. persimilis (fruit fly)	1,375,313	0			–
D. pseudoobscura (fruit fly)	5,161,792	0			–
D. sechellia (fruit fly)	1,203,722	1	+	0/0	+
D. simulans (fruit fly)	2,321,958	7473	+	0/0	+
D. virilis (fruit fly)	3,632,492	0			–
D. willistoni (fruit fly)	2,332,565	2519	–		+
D. yakuba (fruit fly)	2,269,952	0			+
Ixodes scapularis (tick)	13,088,763	44	–		+
N. giraulti (wasp)	540,102	2	+	1/1	+
N. longicornis (wasp)	447,736	1	+	1/1	+
N. vitripennis (wasp)	3,360,694	30	+	4/4	+
Pediculus humanus (head louse)	1,480,551	0			+
Pristonchus pacificus (nematode)	2,292,543	0			–
Tribolium castaneum (beetle)	1,918,906	0			–
GenBank sequence					
Dirofilaria immitis (filarial nematode)	NA	NA	+		+

*This isolate was previously shown to have *Wolbachia* reads in its trace repositories that are contaminating reads from the *D. ananassae* genome sequencing project (*10*).


A
NAV1
H1-hESC
Intergenic DHS
TSS DHS
Gene Body DHS
B
Size of DHS (bases)
4000
3500
3000
2500
2000
1500
1000
500
All DHS
TSS DHS
Gene Body DHS
Intergenic DHS
C
Normalized CG content
1
0.8
0.6
0.4
0.2
0
All DHS
TSS DHS
Gene Body DHS
Intergenic DHS


A
LMOD1
TIMM17A
AoSMC
H1-hESC
HSMMtube
Hepato
Medullo
ProgFib
GM12878
K562
LNCaP
HeLa S3
HepG2
HUVEC
Gliobla
HMEC
MCF-7
Melano
NHEK
Osteobl
Chorion
chr1:201,890,462-201,938,914
# of cell types: 1 1 1 9 2 6 1 18 1 19
B
Cell-Type Specific DHS
Percentage of DHS
0 3 6 9 12 15 18
All DHS
TSS DHS
Gene Body DHS
Intergenic DHS
C
TSS DHS Specificity and CG Content
Normalized CG content
0 0.2 0.4 0.6 0.8 1.0
Open in 1 Cell Line
5
10
15
Open in All Cell Lines