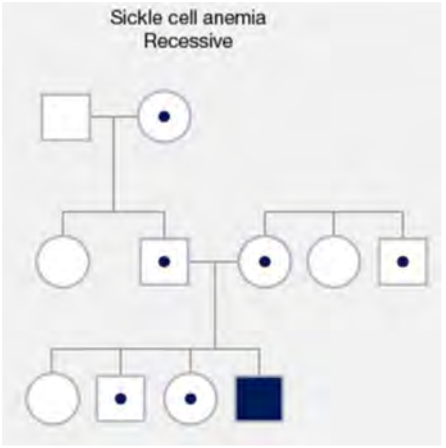
17 slides on the BPRI locust annotation

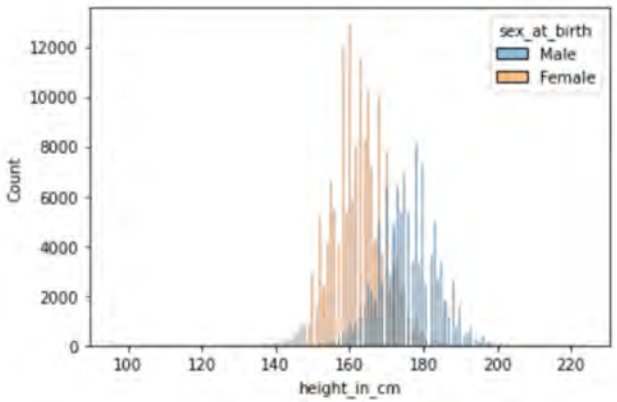
Or: Why the #**$% do we care about genomes and genetics in both humans and life on earth in general?

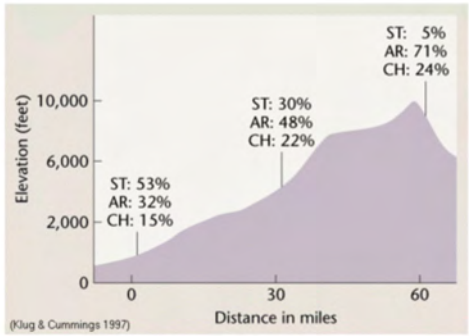
Why do people not really care about genomics?

- “The study of heredity and the variation of inherited characteristics”
- But most people (including many biologists!) fail to understand the scale of its importance for our understanding of life on earth. The study of heredity of form and it’s variation extends to the entire tree of life.
- Yesterday was the 20th anniversary of the human genome release, so It’s part of the furniture, and and everyone takes it for granted. – So since that was a while ago, “What has genetics done for me lately?”
- I think there are multiple reasons for everyone not caring about genetics (beyond the old news), 1. - There are a lot of genes and bases (~~Big~~ Medium Data!) – it’s hard to get your head around what they all do.
- 2. Changes in genotype are read by and acted on by ***cells*** in multiple parts of the body to create changes in phenotype, so the connections can be hard to discern and act on.
- 3. To restate the obvious, evolution doesn’t just select for reproduction to pass on genes, (***survival of the fittest***) but also it selects for plain old ***survival*** even against genetic insult or attack, (The main manifestation of this is diploidy – many of us carry a single copy of genetic mutation that if we had two copies would be cause miscarriage.

Genotype to phenotype at evolutionary scales







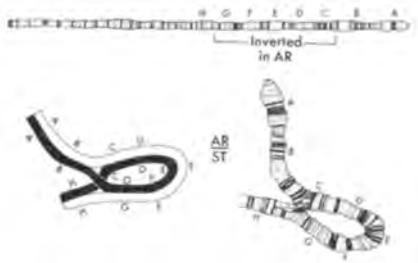
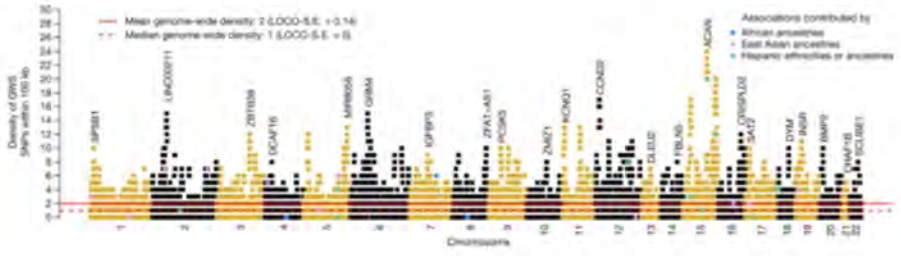


Fig. 2: Brisbane plot showing the genomic density of independent genetic associations with height.

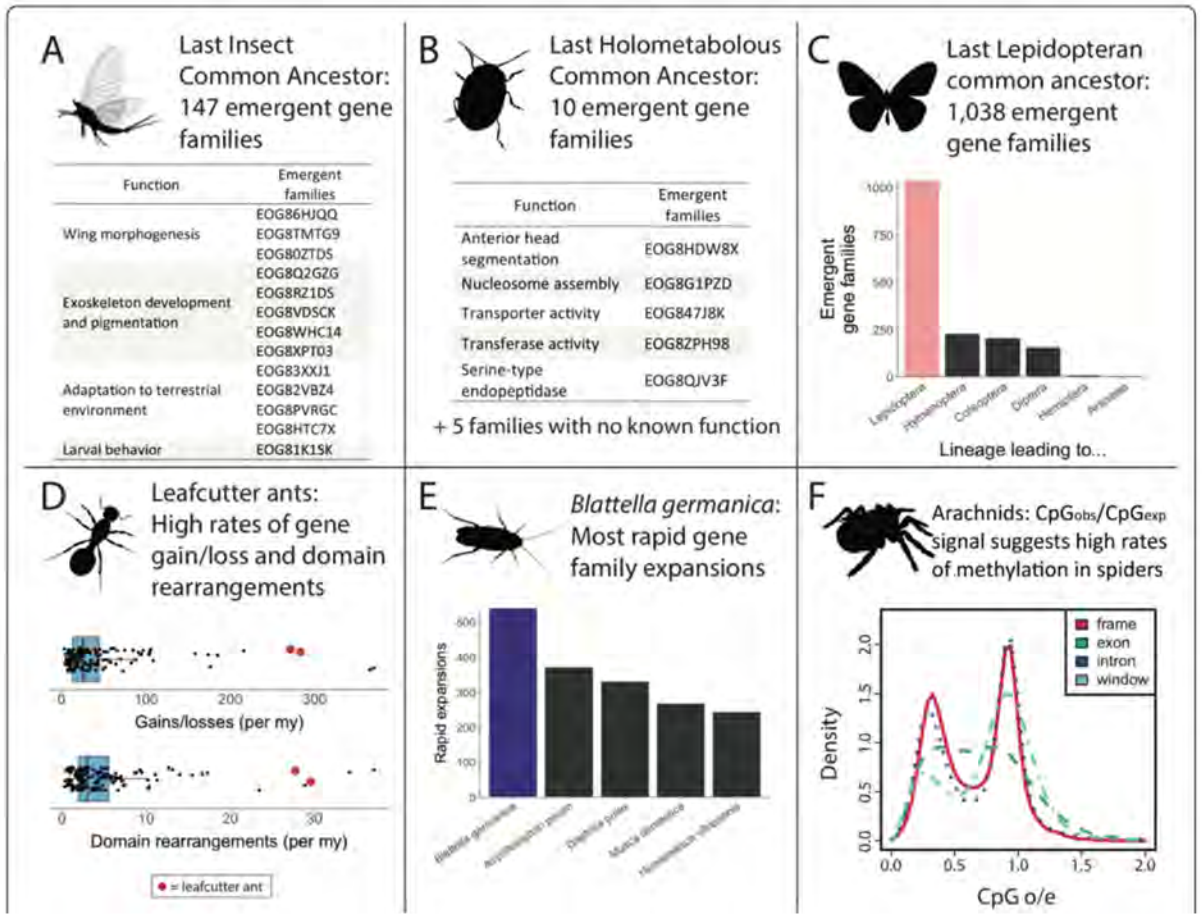


nature › articles › article

Article | Open Access | Published: 12 October 2022

A saturated map of common genetic variants associated with human height

Loic Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh-Huei Chen, Teresa Ferreira,



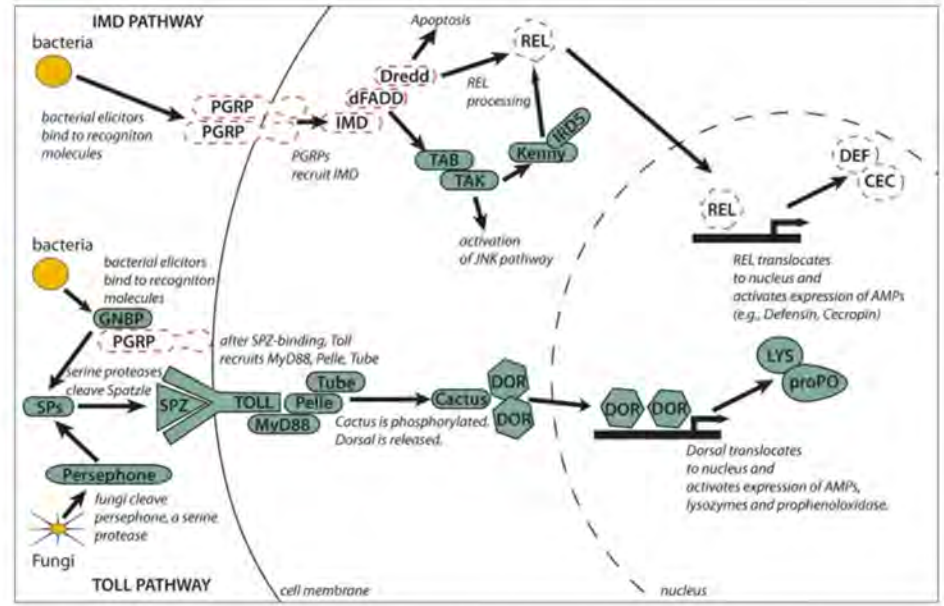


Figure 10. The IMD immune pathway is missing in the pea aphid. Previously sequenced insect genomes (fly, mosquitoes, honeybee, red flour beetle) have indicated that the immune signaling pathways, including IMD and Toll pathways shown here, are conserved across insects. In *Drosophila*, response to many Gram-negative bacteria and some Gram-positive bacteria and fungi relies on the IMD pathway. In aphids, missing IMD pathway genes (dashed lines) include those involved in recognition (PGRPs) and signaling (IMD, dFADD, Dredd, REL). Genes encoding antimicrobial peptides common in other insects, including defensins and cecropins, are also missing. In contrast, we found putative homologs for all genes central to the Toll signaling pathway, which is key to response to bacteria, fungi, and other microbes in *Drosophila*.
doi:10.1371/journal.pbio.1000313.g010

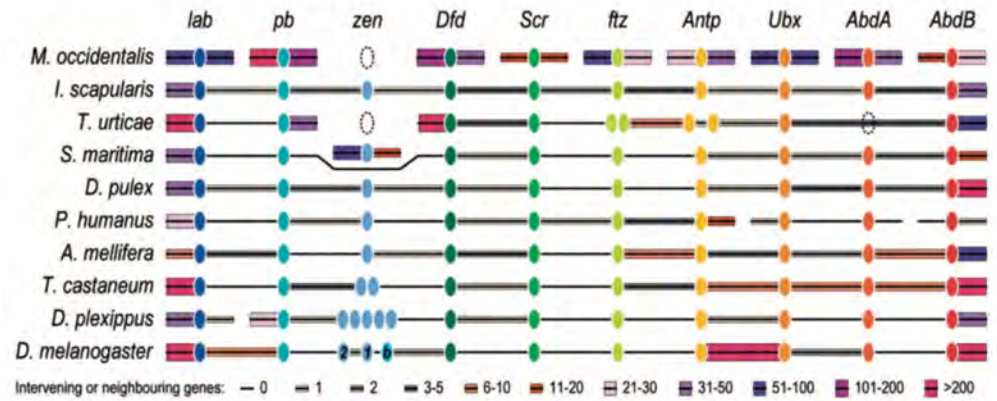
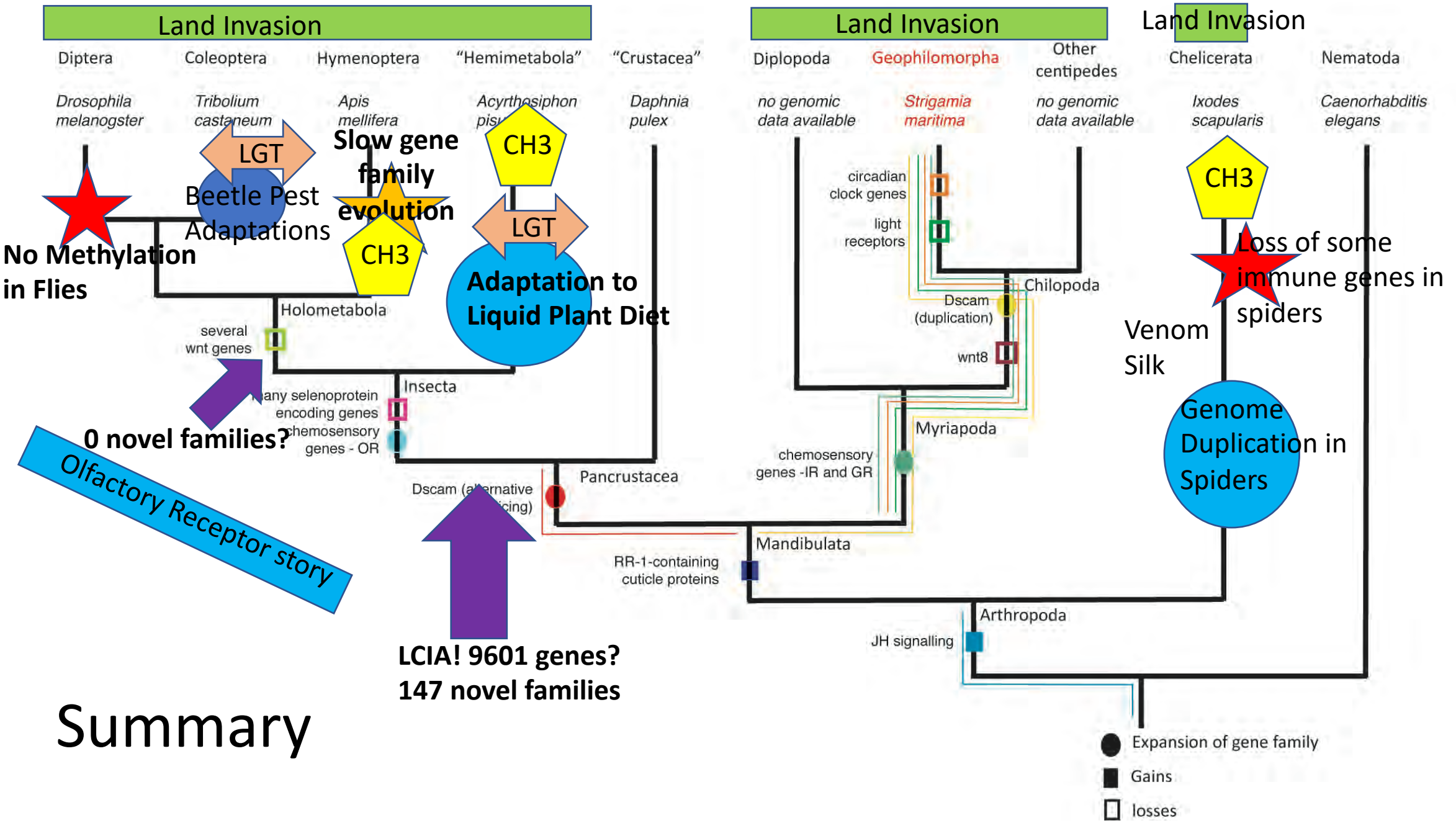
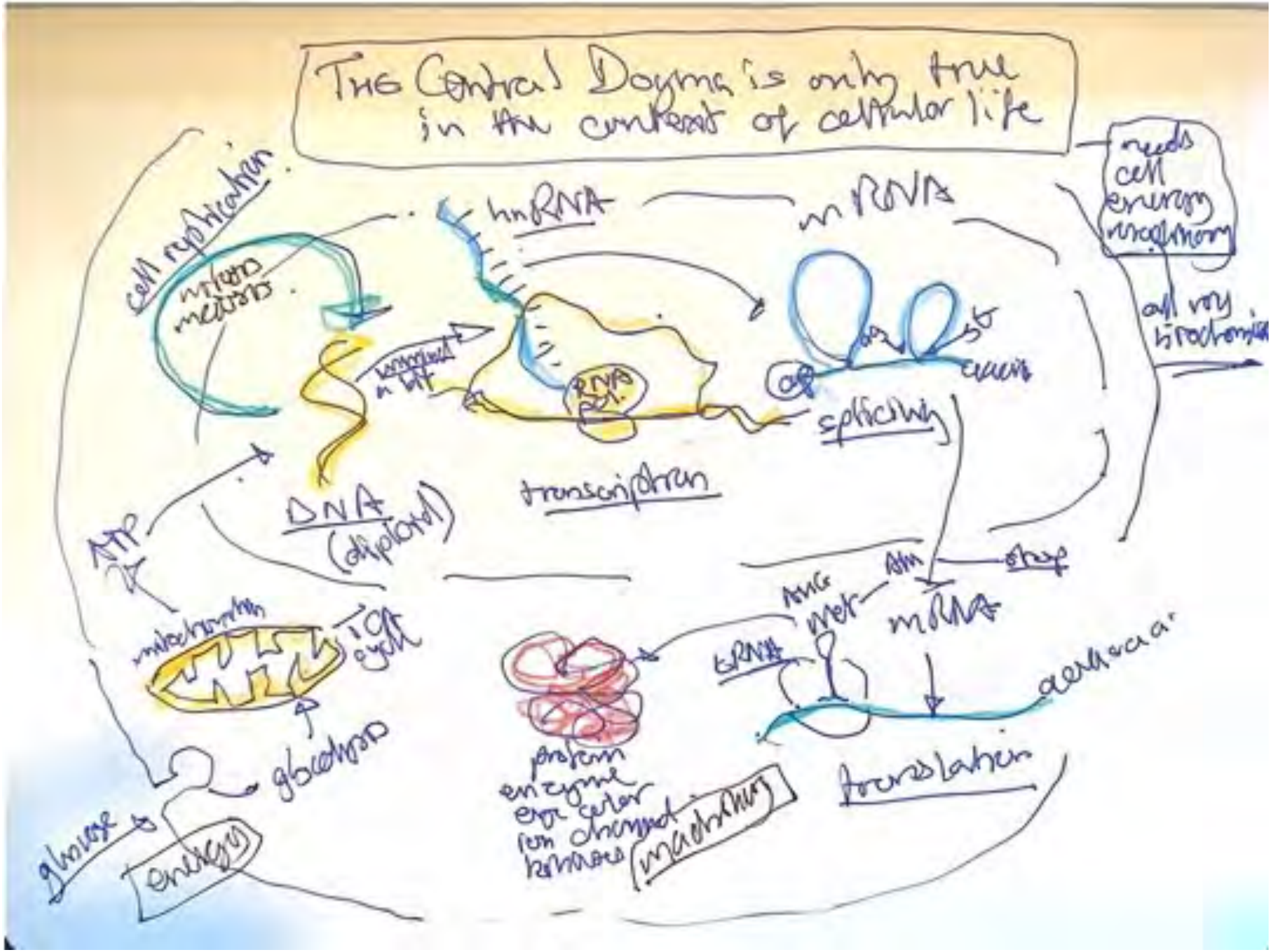


FIG. 3.—Complete atomization of *Metaseiulus occidentalis Hox* genes. In stark contrast to *M. occidentalis*, the genomic organization of ten *Hox* genes


Land Invasion
Land Invasion
Land Invasion
Diptera
Coleoptera
Hymenoptera
"Hemimetabola"
"Crustacea"
Diplopoda
Geophilomorpha
Other centipedes
Chelicerata
Nematoda
Drosophila melanogster
Tribolium castaneum
Apis mellifera
Acyrthosiphon pisu
Daphnia pulex
no genomic data available
Strigamia maritima
no genomic data available
Ixodes scapularis
Caenorhabditis elegans
LGT
Beetle Pest Adaptations
Slow gene family evolution
CH3
CH3
LGT
Adaptation to Liquid Plant Diet
CH3
Loss of some immune genes in spiders
No Methylation in Flies
circadian clock genes
light receptors
Chilopoda
Dscam (duplication)
wnt8
Holometabola
several wnt genes
Venom
Silk
Insecta
many selenoprotein encoding genes
chemosensory genes - OR
0 novel families?
Myriapoda
chemosensory genes -IR and GR
Genome Duplication in Spiders
Olfactory Receptor story
Dscam (alternative cing)
Pancrustacea
Mandibulata
RR-1-containing cuticle proteins
Arthropoda
JH signalling
LCIA! 9601 genes?
147 novel families
Expansion of gene family
Gains
losses

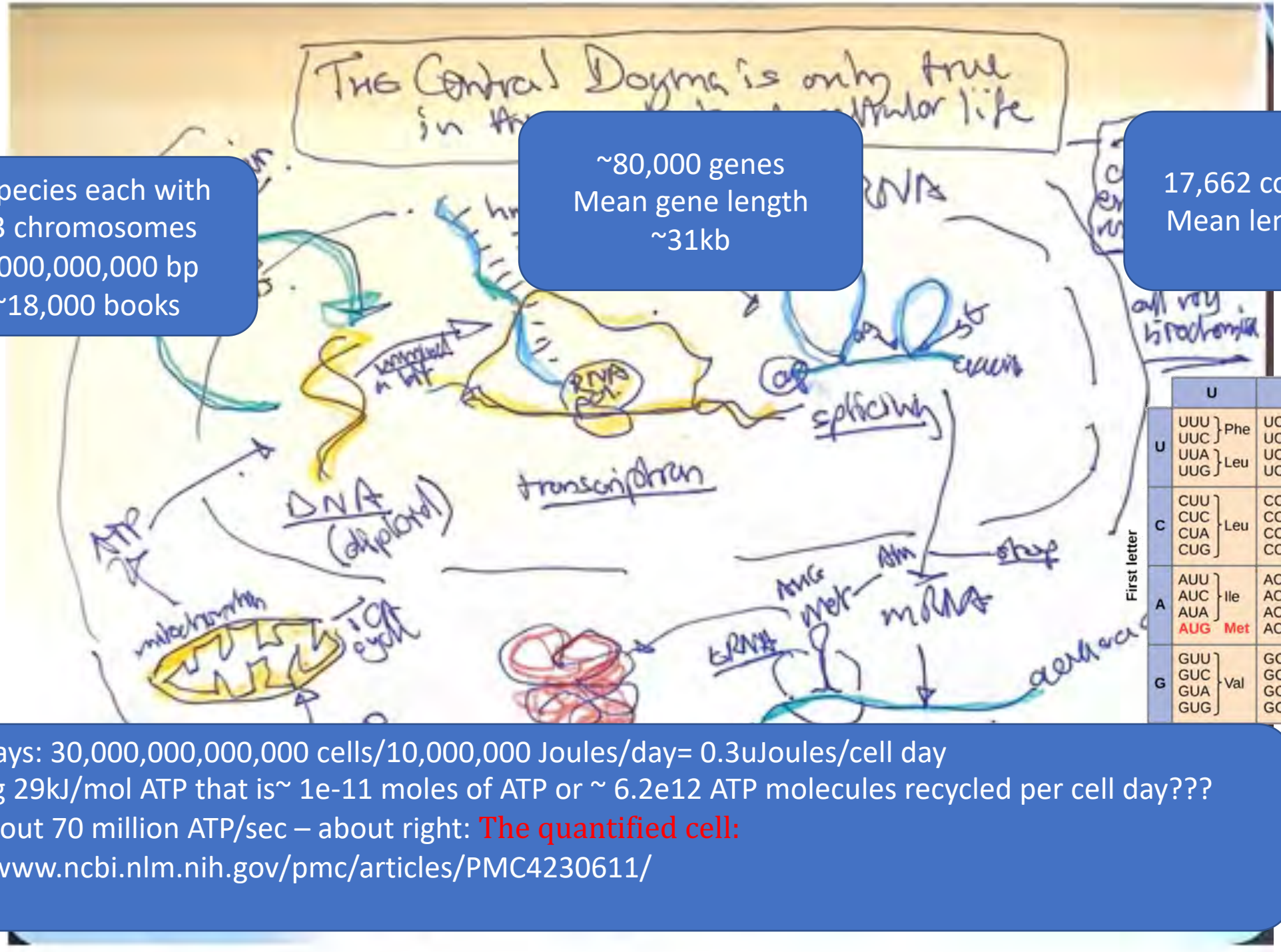
Summary


The Central Dogma is only true in the context of cellular life
needs cell energy machinery
hnRNA
mRNA
cell replication
nucleus
RNA pol.
splicing
transcription
DNA (diploid)
ATP
mitochondrion
glycolysis
glucose
energy
protein enzyme
tRNA
mRNA
translation

6 species each with
13 chromosomes
9,000,000,000 bp
~18,000 books

~80,000 genes
Mean gene length
~31kb

17,662 coding genes
Mean length ~2.5kb



Second letter

First letter	U	C	A	G	
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA UCG Ser	UAU Tyr UAC **UAA Stop** **UAG Stop**	UGU Cys UGC **UGA Stop** UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU His CAC CAA Gln CAG	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU Asp GAC GAA Glu GAG	GGU GGC GGA GGG Gly	U C A G

Google says: 30,000,000,000,000 cells/10,000,000 Joules/day= 0.3uJoules/cell day
Assuming 29kJ/mol ATP that is~ 1e-11 moles of ATP or ~ 6.2e12 ATP molecules recycled per cell day???
This is about 70 million ATP/sec – about right: The quantified cell:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230611/

All the data from NCBI / GenBank.
The difficult part is asking an interesting question.

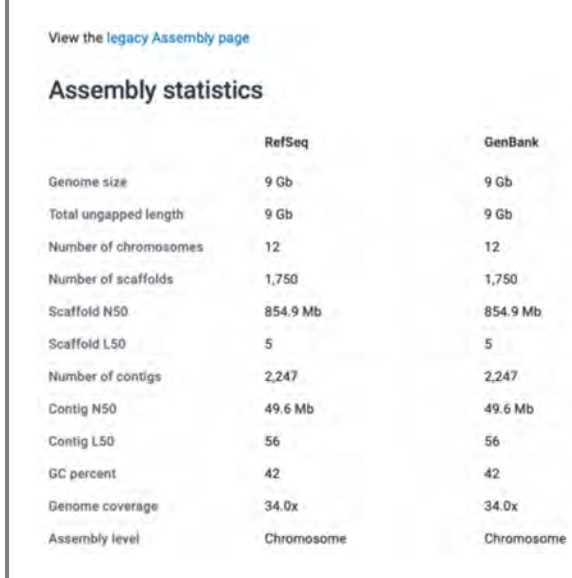
	RefSeq
Provider	NCBI RefSeq
Name	NCBI Annotation Release 100
Date	Mar 3, 2022
Genes	81,274
Protein-coding	17,662
Non-coding	56,385
Software version	9.0

View all genes (includes updated and unannotated genes)

Quality analysis

BUSCO analysis (4.1.4)

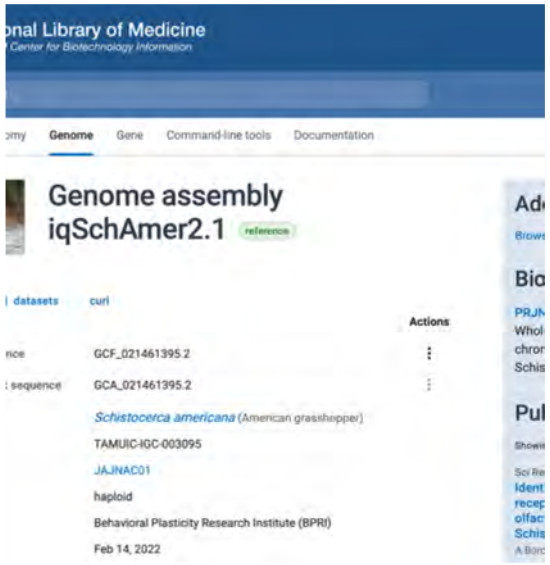
Complete 98.8% (S+D)
Single-copy 96.6%
Duplicated 2.2%
Fragmented 0.1%
Missing 1.2%



View the legacy Assembly page

Assembly statistics

	RefSeq	GenBank
Genome size	9 Gb	9 Gb
Total ungapped length	9 Gb	9 Gb
Number of chromosomes	12	12
Number of scaffolds	1,750	1,750
Scaffold N50	854.9 Mb	854.9 Mb
Scaffold L50	5	5
Number of contigs	2,247	2,247
Contig N50	49.6 Mb	49.6 Mb
Contig L50	56	56
GC percent	42	42
Genome coverage	34.0x	34.0x
Assembly level	Chromosome	Chromosome



Genetics/genomics answerable questions today (by comparison)

- What sex will my child be?
- Who am I related to?
- Do I have a mendelian inherited disease such as Cystic fibrosis like my parents, or what is the chance of passing it to my children?
- What kind of cancer is that tumor we sequenced and what drug is best to treat it?
- Which kinds of microbes are in my gut?
- What kind of virus is in the sewage plant waste water?
- Who left these cells at the crime?
- How many kinds of animals live in this watershed?

Example questions from other species

- My spider is venomous, which are the venom genes?
- Which are the silk genes?
- My mite has no neck, is it missing a hox gene? (yes!)
- Which genes are involved in hearing mating songs?
- My aphid has wing polyphenism, does it have duplicated genes?
- Can we say anything about incomplete metamorphosis, compared with the dragon flies and and holometabola?
- Has there been an ancestral whole genome duplication?
- Is gene expression controlled epigenetically?

Arthropod genomics questions (using comparison)

- What is the population size of this species/population?
- Is there gene flow between these populations?
- What were the historical populations sizes of these species?
- Which genes were lost in the evolution of these species
- Which kinds of microbes/pathogens/symbionts are in their guts?
- Which visual receptors does it have, and what wave lengths can it see?
- Which olfactory receptors let it smell food/water vs mates vs a place to lay eggs?
- How does it defend against transposons?
- How does diet affect it's biochemical capacity? (e.g. carnivores might lose aa synthesis genes, phloem feeders require microbes to synthesize aa's not in their diet)
- How does it make proteins it needs a lot of? (extra copies or more transcription?)
- Does its body plan shape suggest it's missing developmental genes?
- Which genes have different expression between solitary and crowded?
- Which genes have evolved fastest of slowest within the phylogeny?

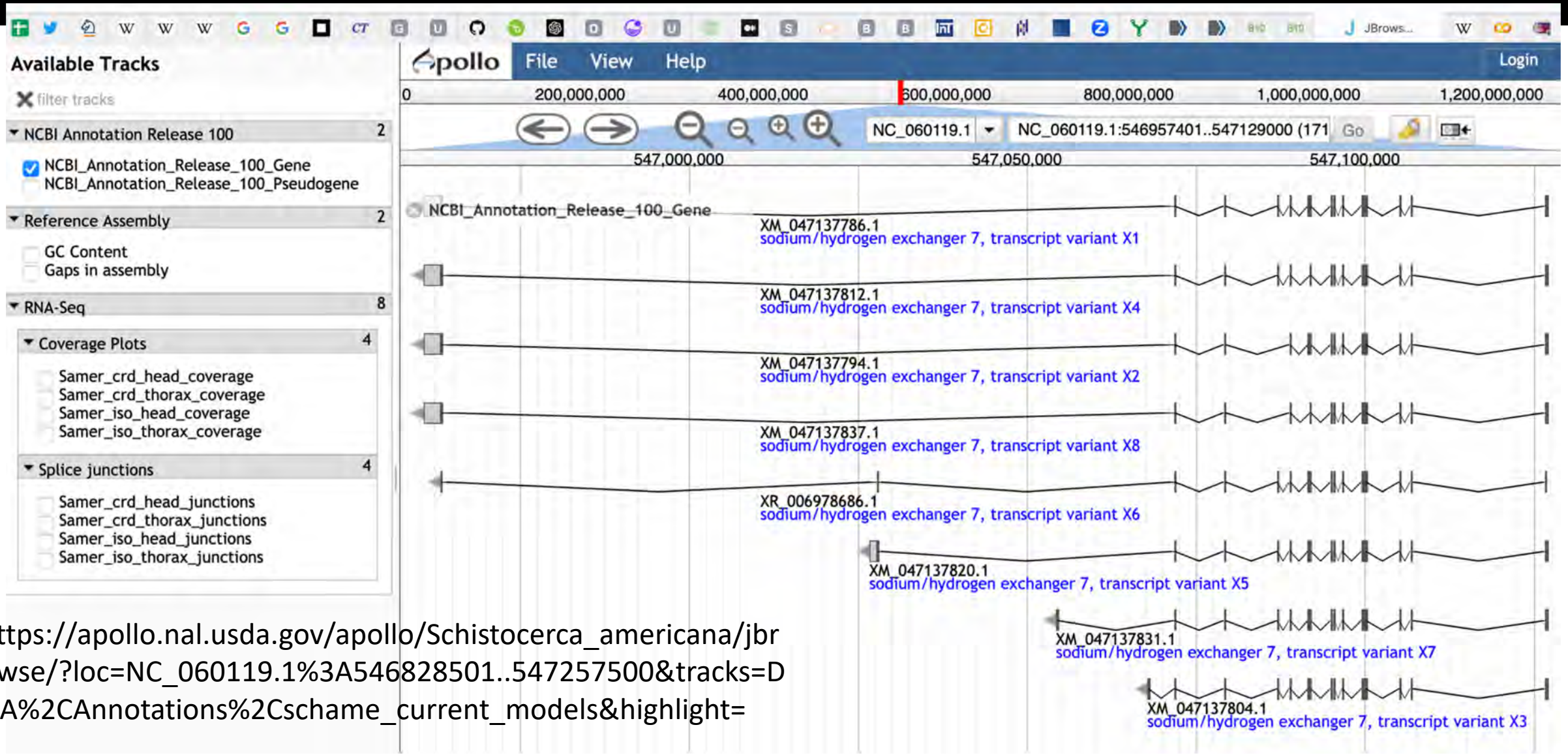
Technical arthropod genomics questions

- How do I design my RNAi experiment guide sequence?
- How do I design my crispr guide sequence experiment?
- Which ion channel is likely effecting my electrophysolgy measurement?
- Does it have one dopamine receptor (or my favorite gene) ortholog, or 2 paralogs?
- Does my favorite gene change between grasshoppers and locust, or does it look the same in all 6 species?
- What protein sequence do I put in alphafold to predict the structure of my protein?
- How inbred is my colony before I need to bring in outside individuals?

Procedure

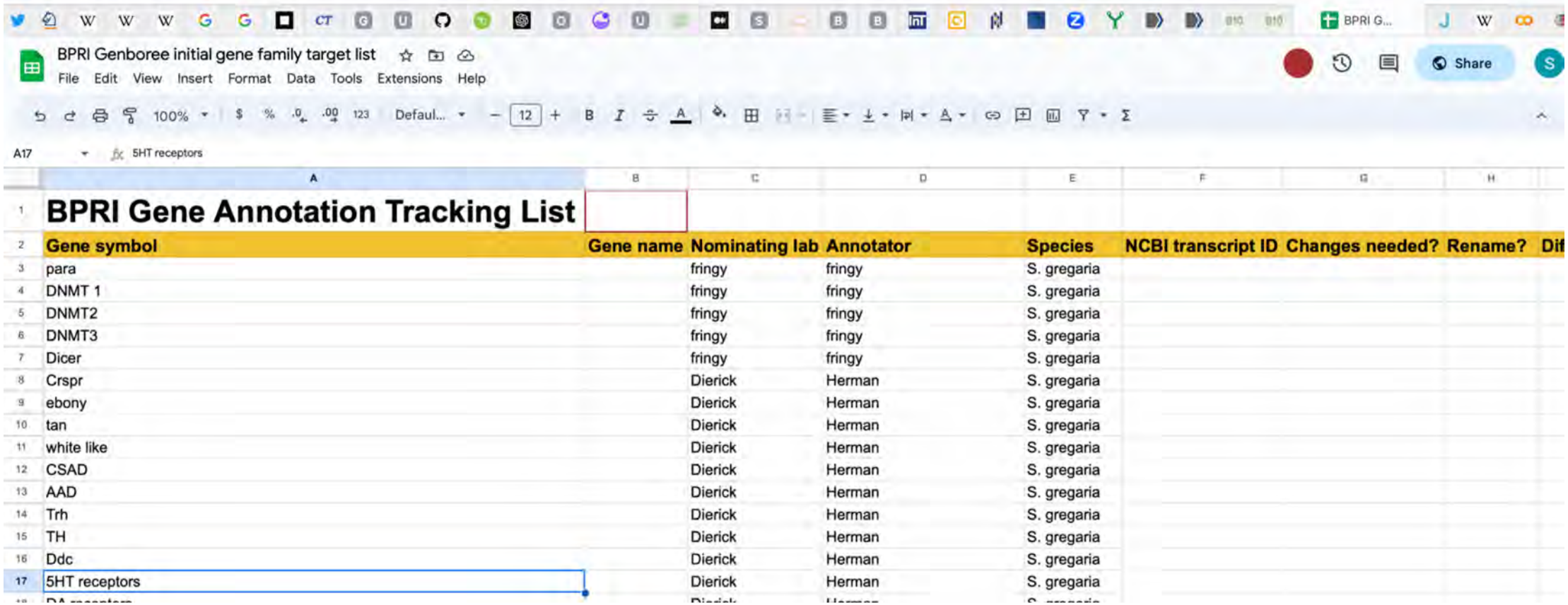
- Participants need to register with the i5k workspace before the Geneboree
- Participants should take advantage of the learning resources before the Geneboree
- We will meet via Zoom on three days from 1 to 5 pm to help guide the process
- It is recommended that participants from each institution gather in a common room to help each other
- Zoom breakout rooms will be created as needed to help participants remotely
- After a tutorial introduction, each participant will be assigned one gene to annotate.
  - The first step will be to do this in Schistocerca gregaria.
  - Next, annotation of the same gene will be carried out in the other 5 species
- An expert will check the work from each participant
- The annotation results will be entered in the Result-tracking Spreadsheet
- Following the initial gene annotation, participants will be assigned additional genes to annotate to the extent possible
- At the end of the Geneboree, each participant will give a lightening talk about his/her results
- Groups of participants working on similar genes will write a summary of their results in one paragraph
- that will be used for future work by BPRI members and to help during the writing of the genome paper.

A random *S. americana* gene:



https://apollo.nal.usda.gov/apollo/Schistocerca_americana/jbrowse/?loc=NC_060119.1%3A546828501..547257500&tracks=DNA%2CAnnotations%2Cschame_current_models&highlight=

Annotation tracking spreadsheet
(please paste in protein sequence when done)



BPRI Genboree initial gene family target list

File Edit View Insert Format Data Tools Extensions Help

Share

A17 5HT receptors

	A	B	C	D	E	F	G	H
1	**BPRI Gene Annotation Tracking List**							
2	**Gene symbol**	**Gene name**	**Nominating lab**	**Annotator**	**Species**	**NCBI transcript ID**	**Changes needed?**	**Rename?**
3	para		fringy	fringy	S. gregaria			
4	DNMT 1		fringy	fringy	S. gregaria			
5	DNMT2		fringy	fringy	S. gregaria			
6	DNMT3		fringy	fringy	S. gregaria			
7	Dicer		fringy	fringy	S. gregaria			
8	Crspr		Dierick	Herman	S. gregaria			
9	ebony		Dierick	Herman	S. gregaria			
10	tan		Dierick	Herman	S. gregaria			
11	white like		Dierick	Herman	S. gregaria			
12	CSAD		Dierick	Herman	S. gregaria			
13	AAD		Dierick	Herman	S. gregaria			
14	Trh		Dierick	Herman	S. gregaria			
15	TH		Dierick	Herman	S. gregaria			
16	Ddc		Dierick	Herman	S. gregaria			
17	5HT receptors		Dierick	Herman	S. gregaria			

Agenda

<u>Rough Agenda Day 1</u> (Wed April 26, 1 – 5 pm)

- 1:00 - 1:30 pm: Introduction to NCBI, Uniprot, i5k Workspace (Hojun)
- 1:30 - 2:15 pm: Gene annotation example (Anna)
- 2:15 - 2:30 pm: Review and assignment of genes to be annotated
- 2:30 - 3:15 pm: Annotate your first gene in S. gregaria
- 3:15 - 4:15 pm: Get back together and show your results
- 4:15 - 5:00 pm: Your next steps

<u>Rough Agenda Day 2</u> (Thu April 27, 1 – 5 pm)

- 1:00 - 1:15 pm: Check-in – answering questions
- 1:15 - 2:00 pm: Work on your next assigned genes
- 2:00 - 2:15 pm: Progress report
- 2:15 - 3:00 pm: Work on your next assigned genes
- 3:00 - 3:15 pm: Progress report
- 3:15 - 4:00 pm: Work on your assigned genes
- 4:00 - 4:15 pm: Progress report
- 4:15 - 4:45 pm: Discussion on anotated genes, tutorial on trees, answering questions (Hojun)
- 4:45 - 5:00 pm: Your next steps

<u>Rough Agenda Day 3</u> (Thu May 4, 1 – 5 pm)

- 1:00 - 1:15 pm: Check-in – answering questions
- 1:15 - 2:00 pm: Work on your genes
- 2:00 - 2:15 pm: Progress report
- 2:15 - 3:00 pm: Work on your genes
- 3:00 - 3:30 pm: Check-in on goals, did everyone annotate one gene?, prepare for lightening talks
- 3:30 - 4:00 lightening talks
- 3:30 - 4:00 Write one report paragraph summarizing the work
- 4:00 - 5:00 Tree generation for a specific annotated gene example (Hojun)

ToDo:

- You must sign up for a web apollo account!!! (like Now!!!)
- i5K workspace registration link:
- https://i5k.nal.usda.gov/web-apollo-registration
- The annotation progress tracking spread sheet is:
- https://docs.google.com/spreadsheets/d/19TUb-_WUREq_iebjmLcQJIOimtDO2WXlfOiyv7Z69sQ/edit#gid=1003083278
- Annotate at least one gene for your labs question!!!
- Schedule: