Busseola fusca Genome Consortium Overview

Using Bioinformatics and Genomics Tools to Develop Integrated, Reproducible, Shareable Workflows

Agenda

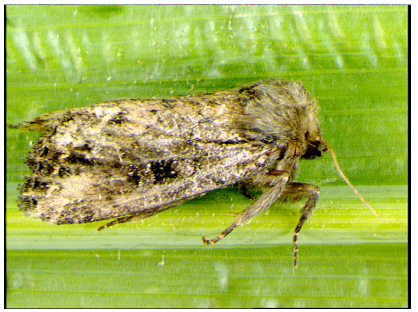
- *Busseola fusca* Genome Consortium

  (tea break)

- The importance of reproducibility and collaboration
  - Data cleaning/ Software version/Parameter choice
  - Record keeping
  - Sharing workflows
- Hands on exercise using Galaxy

Sequencing the Whole Genome and Transcriptome of *Busseola fusca* here at ILRI

Busseola fusca is a major crop pest, especially of maize



B. Fusca devastates large proportions of both small & large-scale crops in sub-Saharan Africa

TABLE 1. Mean percentages of plants infested with eggs (±SE), number of egg batches per plant, discovery efficiency and percentage egg parasitism on *Busseola fusca* at the whorl stage of maize in four agroecological zones (AEZ) in Uganda during 2004[z]

Season	AEZ	Egg batches/plant	Discovery efficiency (%)	% Egg parasitism
1st rains 2004	Eastern	0.010±0.005 a	9.8±6.0 a	39.4±16.9 a
	Southeastern	0.016±0.005 a	9.6±4.8 a	17.7±8.4 a
	Lake Albert Crescent	0.00 b	NF	NF
	Lake Victoria Crescent	–	–	–
	df	2, 62	1, 47	1, 19
	F value	4.11	1.05	1.32
	P value	0.0211	0.3465	0.2642
	Eastern	0.007±0.003 a	5.6±5.6 ab	27.2±12.4 ab
	Southeastern	0.020±0.005 a	18.8±7.4 a	46.1±10.2 a
	L. Albert Crescent	0.015±0.006 a	1.8±1.8 ab	7.4±5.0 b
	L. Victoria Crescent	0.004±0.002 a	0.0±0.0 a	0.0±0.0 b
	df	3, 73	3, 73	3, 30
	F value	2.48	2.99	3.90
	P value	0.0680	0.0365	0.0183

means followed by a common letter do not differ significantly at $P \leq 0.05$ (SNK).

ed.

NF, host not found.

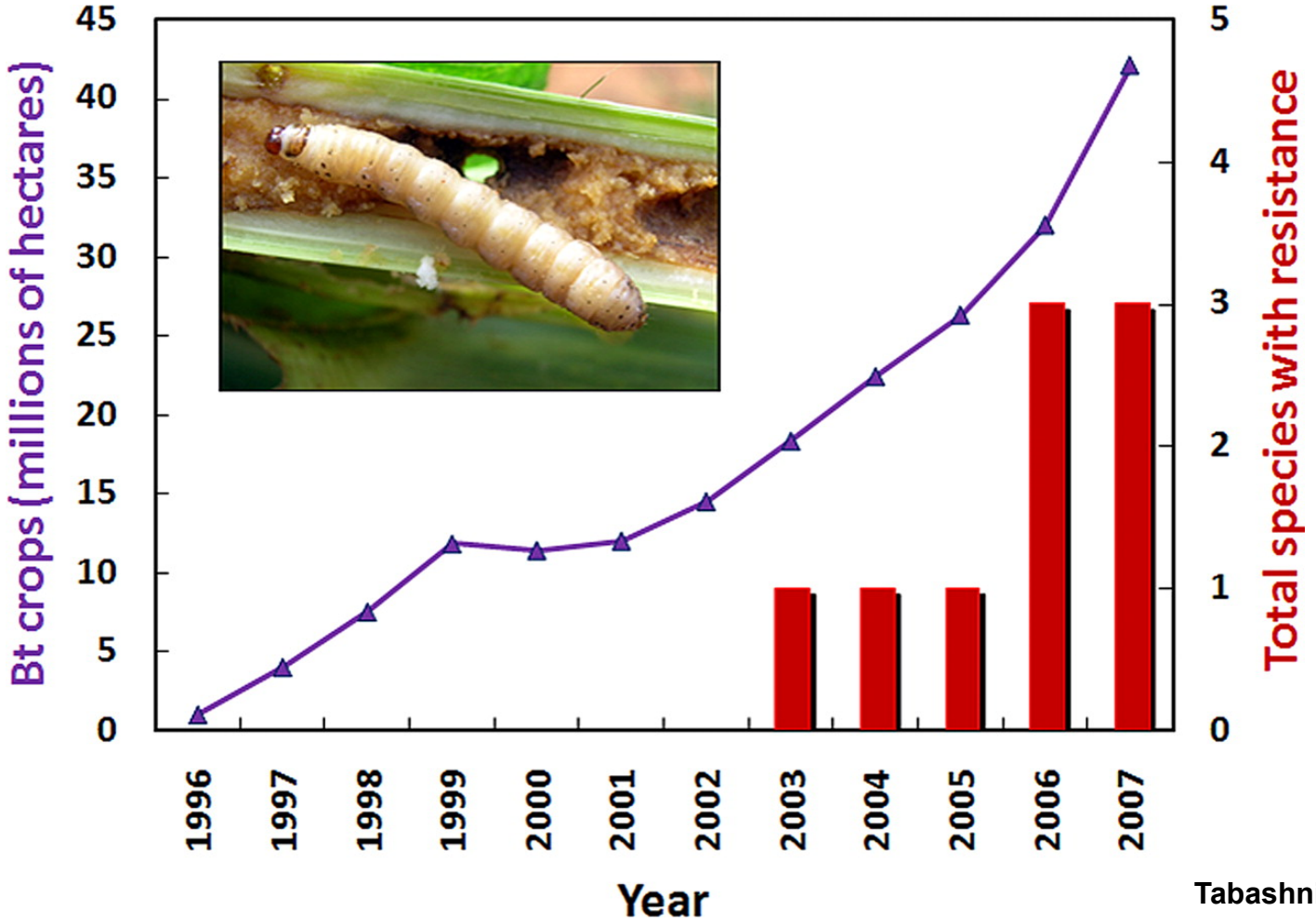
Matama-Kauma et al. 2007

Current Methods of Pest Management

Cultural control – destroying crop residues, intercropping, rotation, etc.

Synthetic pheromone traps – disrupt mating

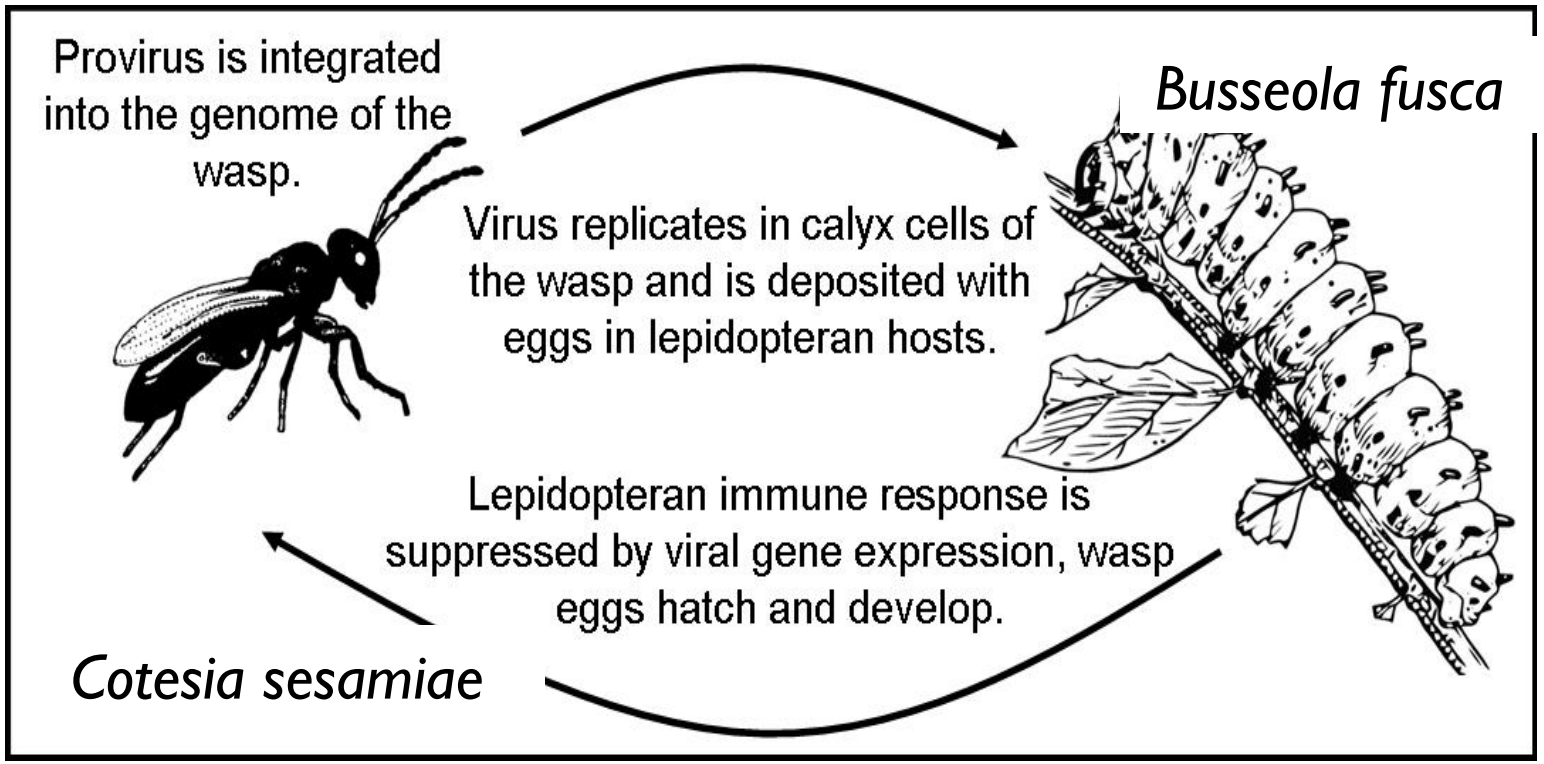
Breed resistant crops – but larvae evolve resistance to the resistance!



Tabashnik et al. 2008

Biological Control for Pest Management

Parasitoid wasps use *B. fusca* larvae to lay eggs,
and simultaneously deposit viral particles
(the virus is embedded in the wasp's genome).

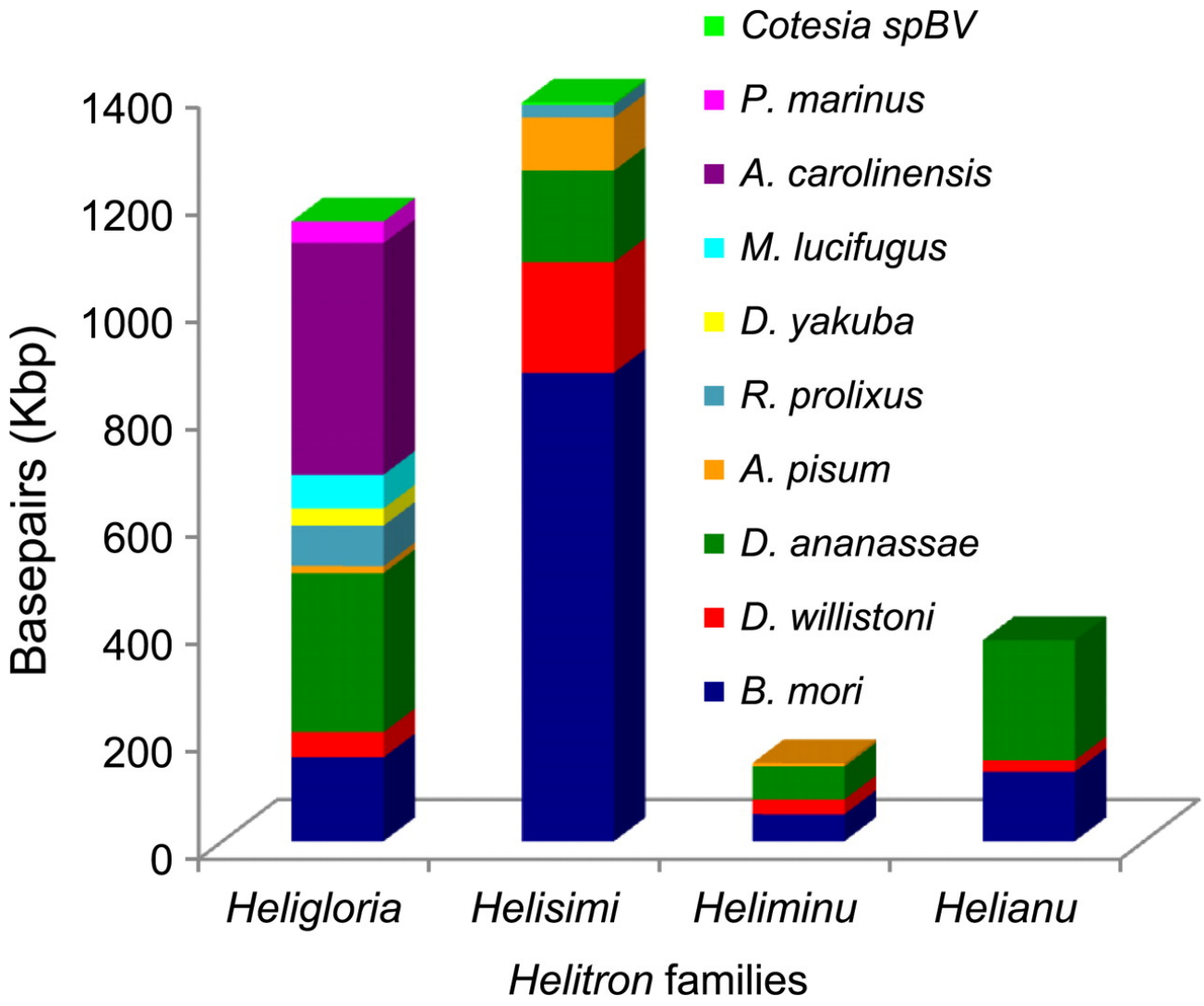


Fewer chemicals, affordable, widespread/sustainable, efficient.

Multiple Reasons to Sequence *B. fusca*

- ***Major* crop pest threatening food security in East Africa**

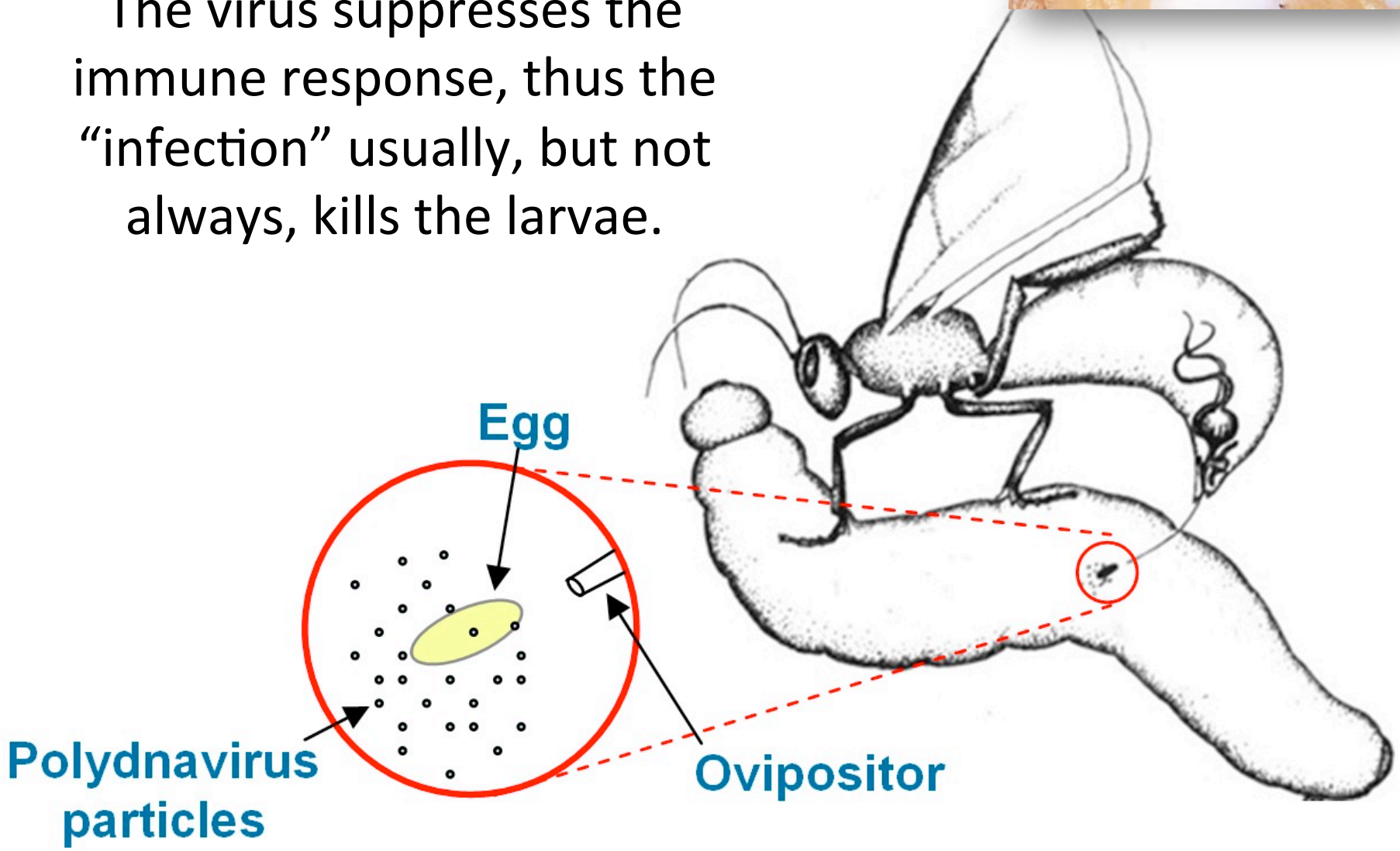
- Previous cases of HTT involve the virus in the parasitoid wasp of *B. fusca*



Thomas, Schaack, & Pritham 2010

Closely-Associated Species Can Exchange DNA Horizontally

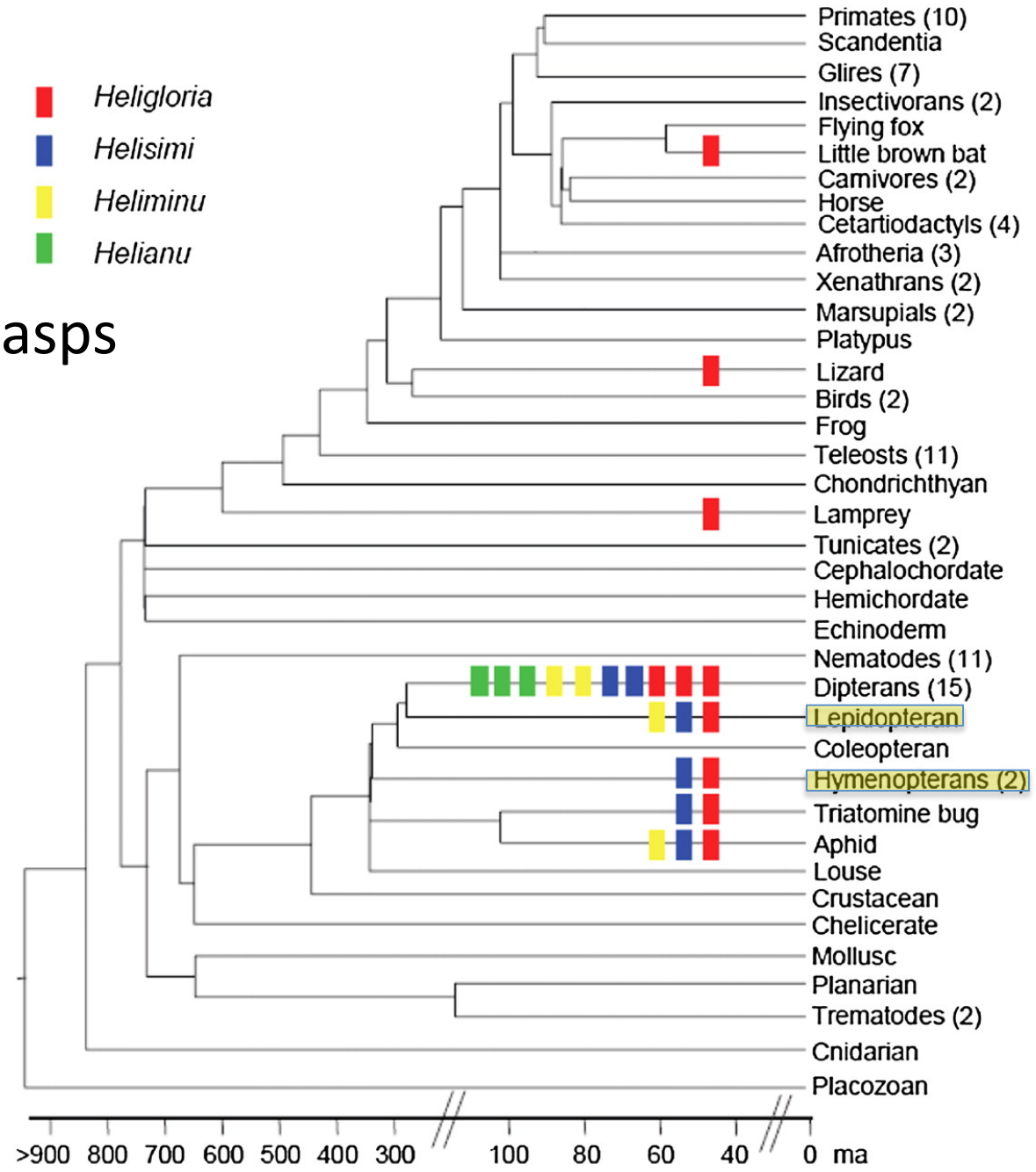
The virus suppresses the immune response, thus the "infection" usually, but not always, kills the larvae.



Nearly identical *transposable elements* have been found in the WGS of various organisms *including*:

the viruses in *Cotesia* wasps and the first sequenced lepidopteran, *Bombyx mori*

A TE is a piece of DNA that is, or once was, capable of moving or replicating and reinserting in the genome.



Thomas, Schaack, & Pritham 2010

Unexpectedly, *many* cases of horizontal transfer among eukaryotes found in WGS data

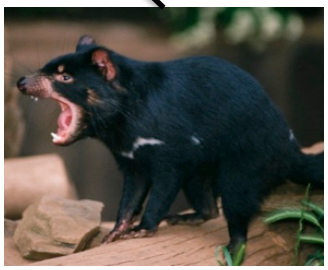
Many different types of TEs

Many species (>13)

Spanning 4 phyla

On 4 continents

But how?

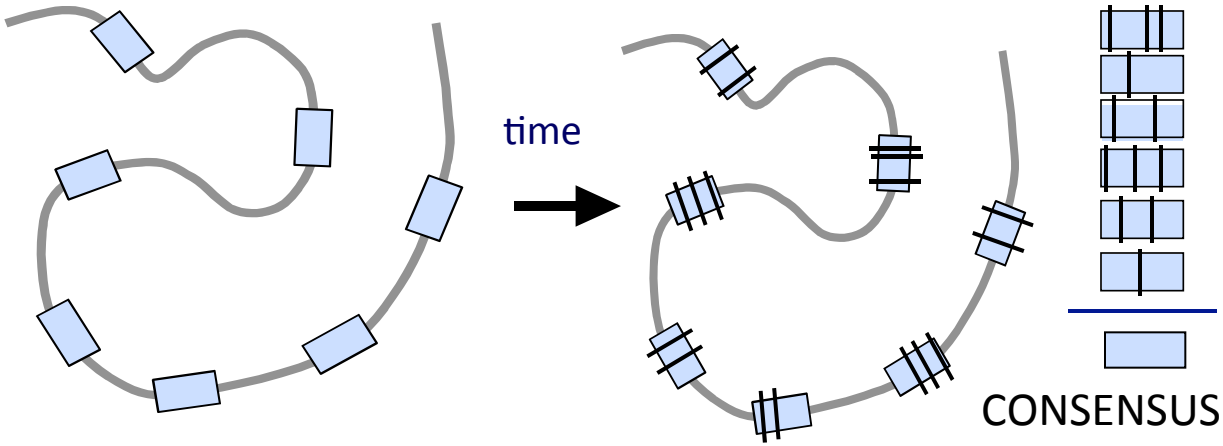


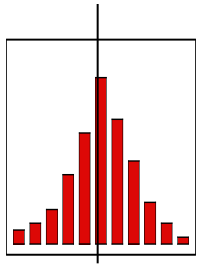
Gilbert/Schaack et al. 2010
Schaack/Gilbert et al. 2010
Gilbert et al. and Schaack 2013

Searching WGS Data for Evidence of HTT

>contig00001 length=577 numreads=6
TGCGGGACTCGGTAGACGAATTGCTCAGGTGCTTCTCGCTCCTGCTCGGAAaCTttGTTC
TCAGCCTCGTAATTTCTCTGATTCTGGTACTTGTTATCGTCGAACTGCGCCGGGCCGTAG
TTTTCGTCGGCTGAAGACTTGACTTGCTGGTGATGATTGTTGGGGTGGTAGACCAGCTTC
GGGCGTTCCATCCTGACGCGTTGGATAATTTTCGGGTtGTtCGTCGTGGTGAACAGCGGA
CGAGGAGTCGAAAGTGACCCTGCGTGGTAGACGTCAGAGTGTTCGTCGCTGCTTTCTTGG
GATTCGCGAACCTTGATGGTAGTTTCGTCTTGCAGCTGCCGGCGTAAGTTATTGCCACTG
ACCGGCGACTTGGACTCTTCTTcTTTTTTCCTCAAACTCTTCAGTTATTGGACGGAGGGTG
TGATCAAGCGGCCGTTTTGTGAAGTAGCTCTTACGCTGCGTGGCGGCTTTGGCTTCGACG
GTTGGCCGAGTGATCTGCGAGTCTTCTTCCAAGTTGTCCGGCGATTCTTGCTCACTGGAC
TCTGGCTCCGGGCTTGATCTCaagTTTGTGAACAaTC
>contig00002 length=3271 numreads=157
aaTAgAAtCtAATTTTTATATAAaTCACTTTATTcTTATTaTTAACTTGATGAAAACTTT
TTtCTTCGTGTTAAAACATATATTTAGATTCCTGACCTGTTATTGTTTCAAACTTAGTTT
GGCTTCAGCTTACTTATAGATGCTTTAAGTTACCTACTTCATTTTAGTGTGACTCCACCG
TAGATATCGAGCCAAAAATTCGTATCTGTGTAACGCTTAAATAGAAACATCAAGTCGAAT
TTGTCGATATCTATTCATCTATCAGTAAGTCTTACAACGTGGATTCATTCTGTTCTTATT
TTCAACAAAAAGTGTATTTTTTTTTcAaTGGTCCGAGCGAATCCGCAAGAAGTAATGCAG
TCTGCATAACGTTGAATTTATGAGCAACTAACCTGCTCTGTTTTGCAAGGACTAGAATCA
GTTGCATTGAGCAACATTGCTTTCAAGTCAGTAGAGTCTTCTTGCAATATAAGCGCTAAA
AAAGTTTTCTTGAACAACATTAGAGTTCCGAGTTCCCAACAAGCTTTCTAAATAACTGGA
ACGGAAGTTGTTCTAACTAACAACGAATTCGTCGGACTGGAACAGGAACATTACATAGTG
GGCCATAAAGTGCATCTAGAAAATAATACATATCACGGCCAACTCCAACTTCATCAAACA
GCAGGAAATGAAATTGACATAGTAAGAAATATCTATTATGGTCAAGTCCAAATCCATGAG
TTACTCGTAACTCAAGTAAAACgaTCGTTCAATACTTATAATAAACTTCCAAATAACTTA
CCTATTGGAAACTGTGCAAGATTTATTCAAAGTGTGTATAATAAATCGAAGGCTGGAAGT
GAGGTAGCTTTAAGGCTCGTTAGAAAAATTACTCGCAAGTGCAGTACTTTTTCTTCGGAA
ATGTAAATAACAACACTTAGTCGTAATTTTTTTTTTTtGAatAACTTTTATTGTAaCTA

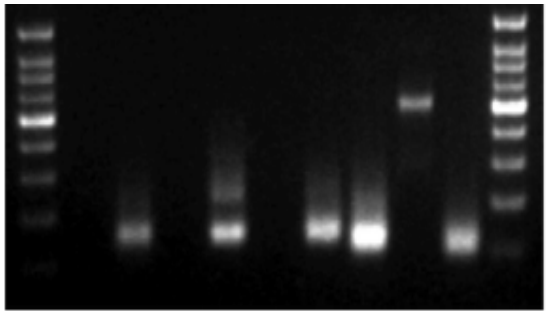
Searching WGS Data for Evidence of HTT

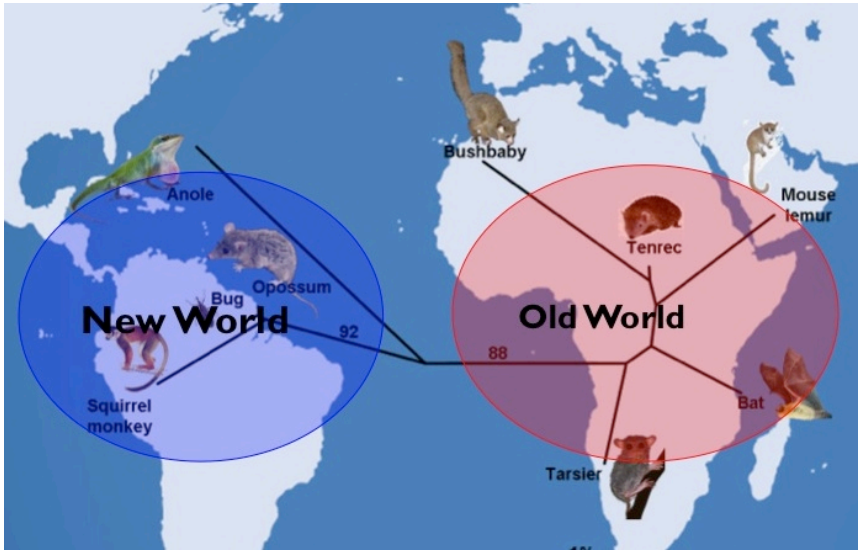




Pairwise divergence

PCR/
Cloning/
Sequencing
to validate

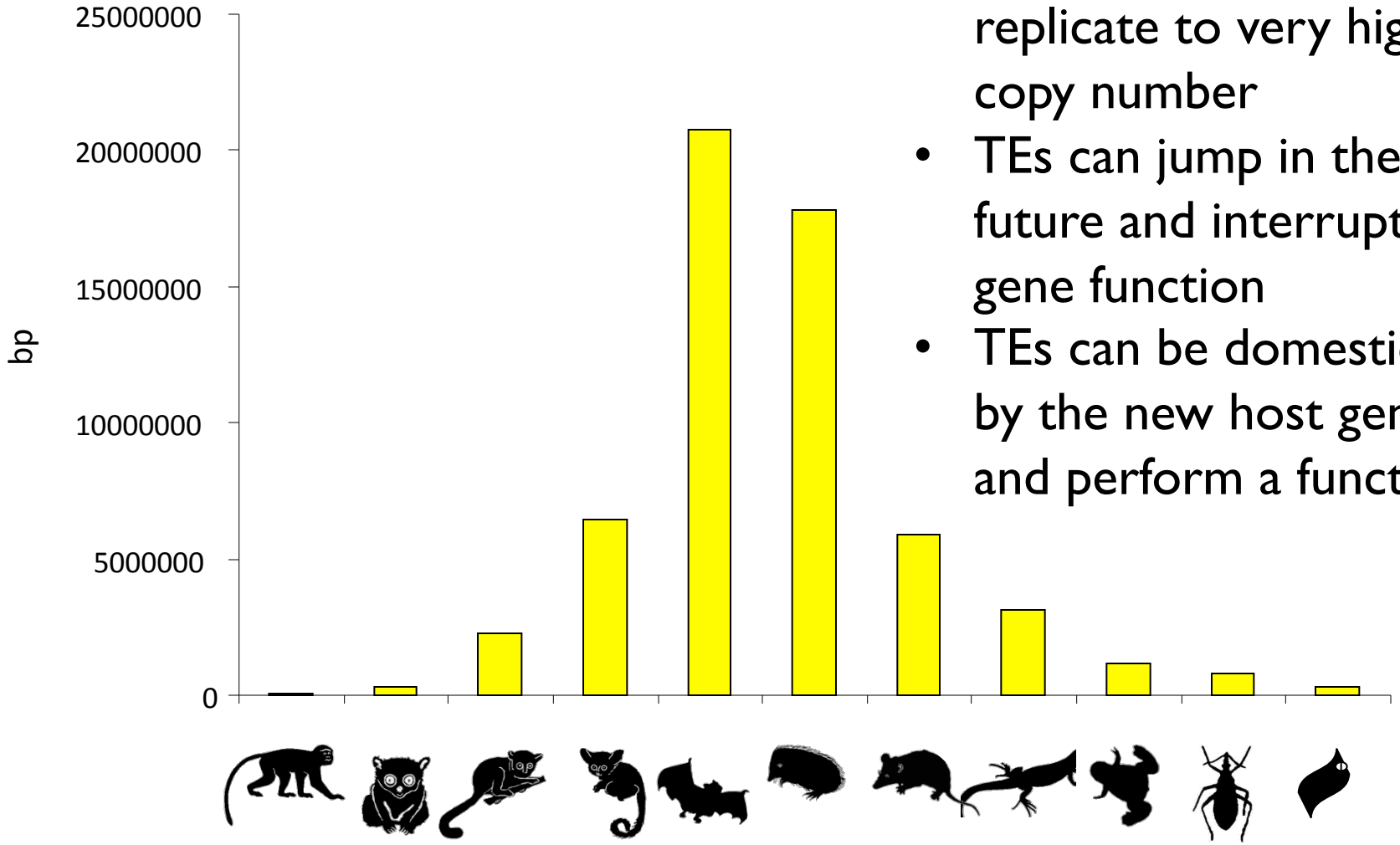




Biogeographical
& natural history data

Average divergence X Mutation Rate = Calculate the date of invasion

But who cares if TEs are exchanged horizontally?



- After HT, TEs can replicate to very high copy number
- TEs can jump in the future and interrupt host gene function
- TEs can be domesticated by the new host genome and perform a function

In bacteria, we knew HTT was frequent and important.

In eukaryotes,

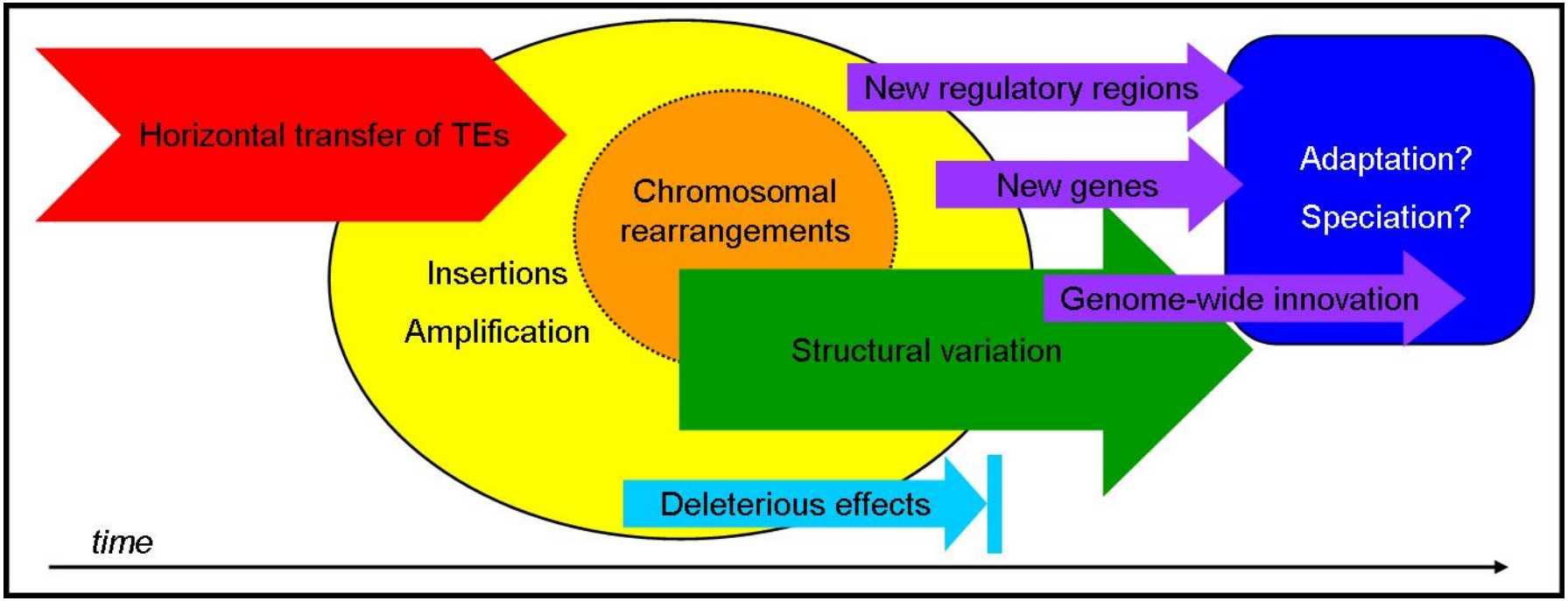
we now know that HTT is not uncommon.

What are the mechanisms?

How are TEs incorporated into the germline?

What is the impact?

How often does this happen?



WGS and Transcriptome Sequencing

Genome sequencing:

DNA extracted from a single larva (1 run, ~15x coverage of 900 MB)

WGS and Transcriptome Sequencing

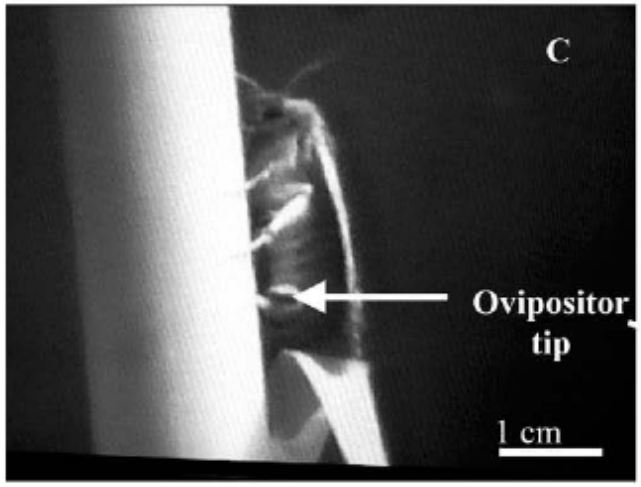
Genome sequencing:

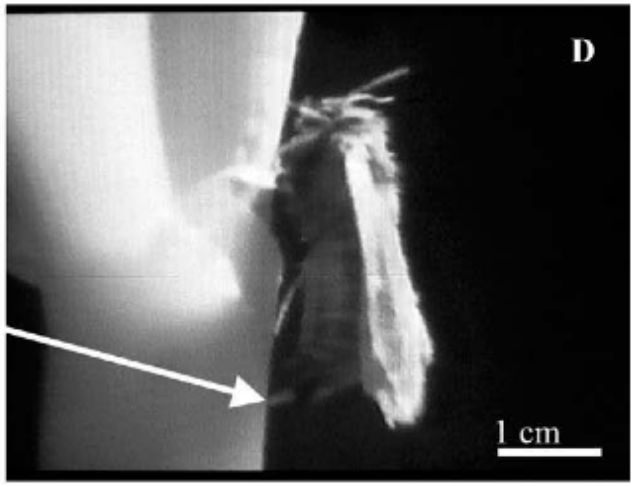
DNA extracted from a single larva (2 runs, ~30x coverage)

Transcriptome sequencing:

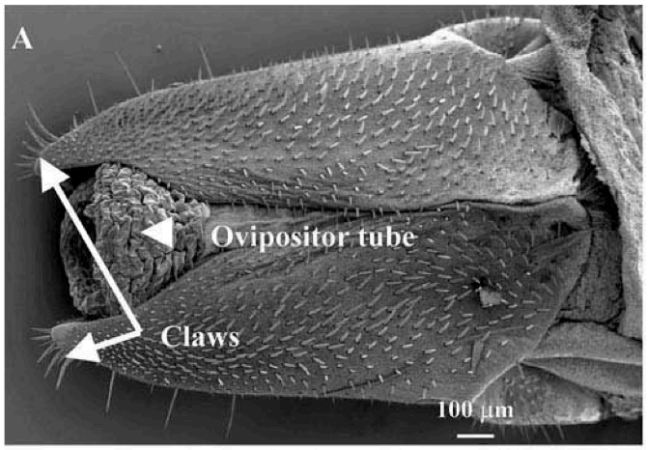
S2_S1_L001_R1_001.fastq	B. fusca unexposed; larvae (1)
S3_S2_L001_R1_001.fastq	B. fusca exposed to CSM; larvae (1)
S6_S3_L001_R1_001.fastq	B. fusca female thorax (1)
S7_S4_L001_R1_001.fastq	B. fusca male thorax (1)
S10_S5_L001_R1_001.fastq	B. fusca female ovipositor (1)
1_S1_L001_R1_001.fastq	B. fusca neonates (350)
2_S2_L001_R1_001.fastq	B. fusca eggs (700)
3_S3_L001_R1_001.fastq	B. fusca male antennae (60)
4_S4_L001_R1_001.fastq	B. fusca female antennae (60)
5_S5_L001_R1_001.fastq	B. fusca exposed to CSK; larvae (1)

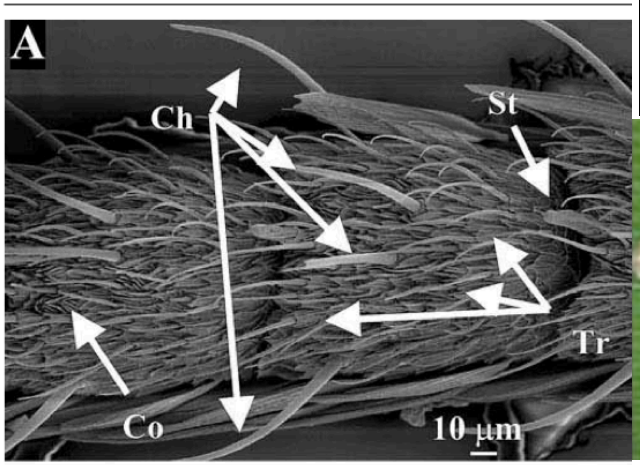
Cataluyud et al. 2006





Genes upregulated in different developmental stages, organs, males versus females, and exposed and unexposed to wasps





Differential expression of the CrV1 haemocyte inactivation-associat polydnavirus gene in the African maize stem borer *Busseola fusca* (Fuller) parasitized by two biotypes the endoparasitoid *Cotesia sesamiae* (Camer

C.W. Gitau[a,c,*], D. Gundersen-Rindal[b], M. Pedroni[b], P.J. [illegible] Dupas[d]

entre of Insect Physiology and Ecology, P.O. Box 307[illegible] Kenya
A-ARS Insect Biocontrol Laboratory, Beltsville, M[illegible]
Kenyatta University, P.O. Box 43844-00100, N[illegible]
e Populations Génétique et Evolution, 1 av[illegible]198 Gif-sur-Yvette, France

mber 2006; received in revised form[illegible] cepted 11 April 2007

Ab

heir na[illegible] immune suppressive abilities. The polydnavirus harboring braconid *Co*[illegible]sito[illegible]*sca* and *Sesamia calamistis* in sub-Saharan Africa exists as two biotypes. In Kenya, the western biotype completes de[illegible] *fusca* larvae. However, eggs of the coastal *C. sesamiae* are encapsulated in this host and ultimately, no parasitoids e[illegible]sitized *B. fusca* larvae. Both biotypes develop successfully in *S. calamistis* larvae. Encapsulation activity by *B. fus*[illegible]s eggs of the avirulent *C. sesamiae* was detectable six hours post-parasitization. The differences in encapsulation [illegible]virulent strains were associated with differences in nucleotide sequences and expression of a CrV1 polydnavirus (P[illegible] is associated with haemocyte inactivation in the C[illegible]V1 expression was faint[illegible] body and haemolymph samples from *B. fusca* parasit[illegible]ich exhibited encaps[illegible]Expression was high in fat body and haemolymph samples fr[illegible]vae parasitized b[illegible]*sesamiae*, encapsulation in the former peaking at the same time p[illegible]on synonym[illegible] CrV1 gene sequences between virulent and avirulent wasp suggests[illegible]by *C. se*[illegible]ue to qualitative differences in CrV1-haemocyte interactions.

Cotesia sesamiae Mombasa (CsM) causes an immune response and can't infect *B. fusca* successfully

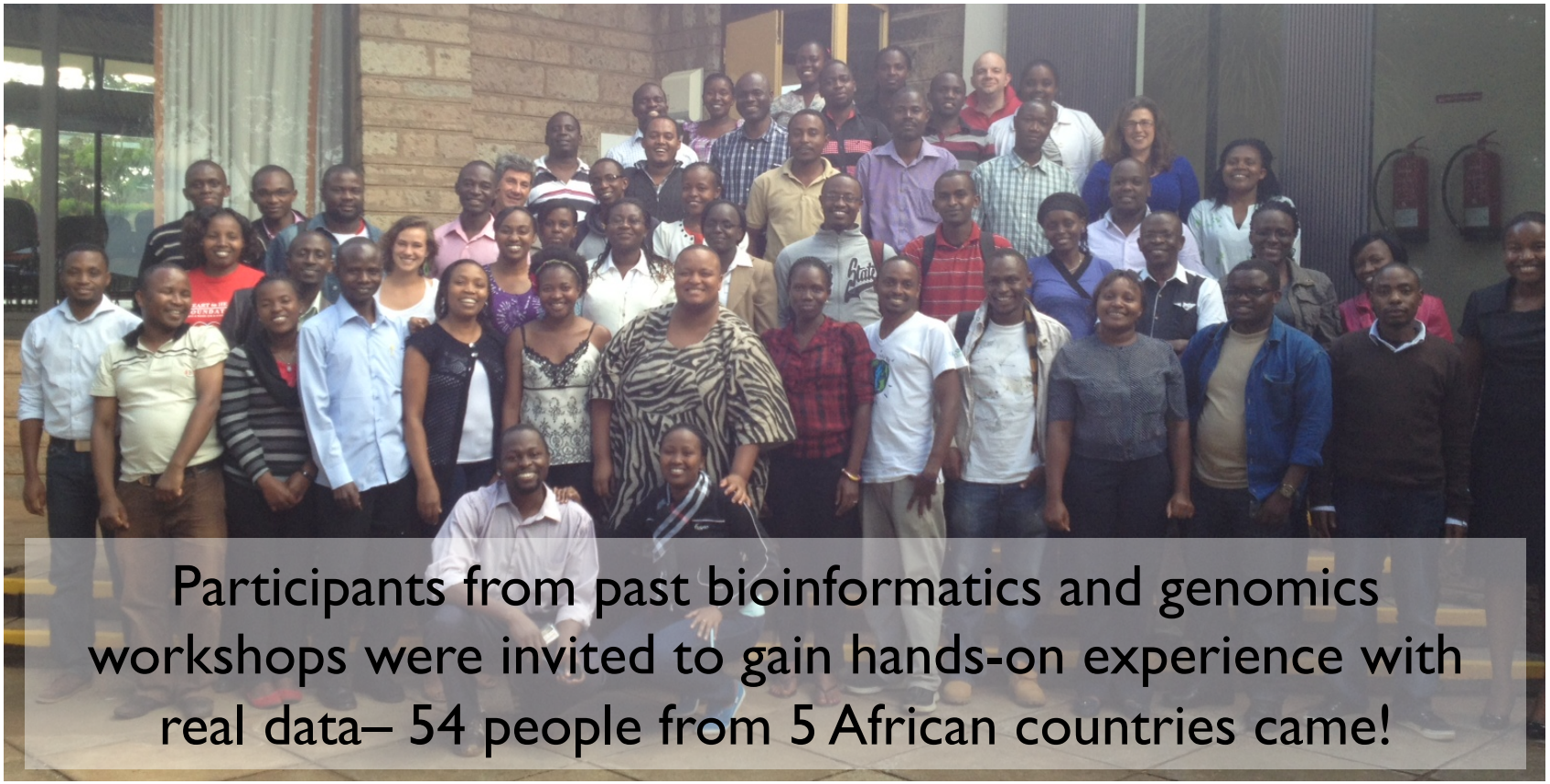
Cotesia sesamiae, a widespread endoparasitoid of *Busseola fusca* and *Sesamia calamistis* in sub-Saharan Africa exists as two biotypes. In Kenya, the western biotype completes development in *B. fusca* larvae. However, eggs of the coastal *C. sesamiae* are encapsulated in this host and ultimately, no parasitoids emerge from parasitized *B. fusca* larvae. Both biotypes develop successfully in *S. calamistis* larvae.

Cotesia sesamiae Kitale (CsK) CAN infect *B. fusca*

Genomics and Transcriptomics of *B. fusca*

integrating research and capacity building

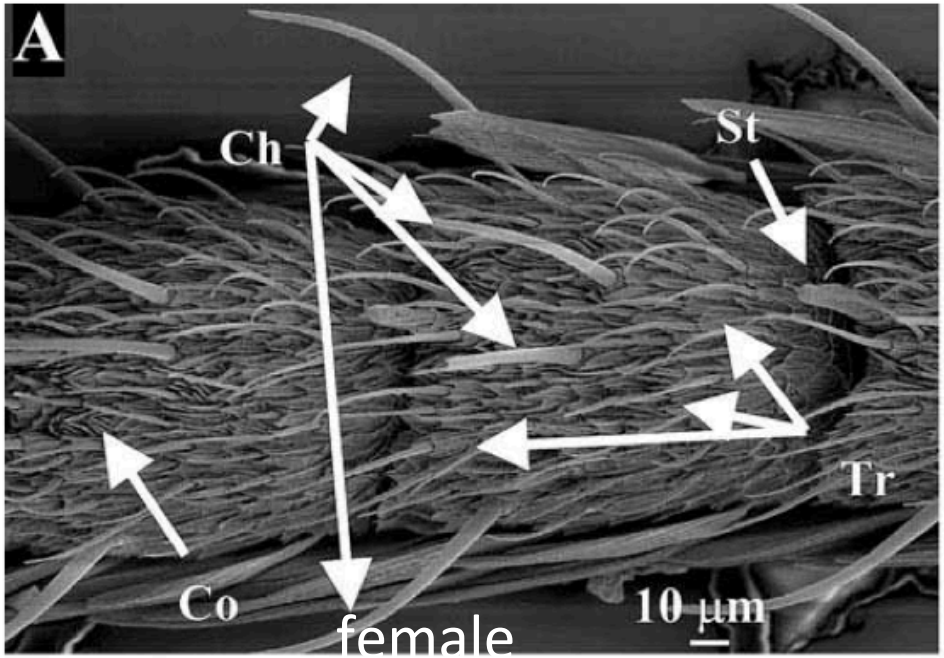
- Animals reared at icipe and sequencing at ILRI
- Transcriptomes from 10 different tissues
- Assembly and annotation jamboree in Nairobi 2 weeks ago

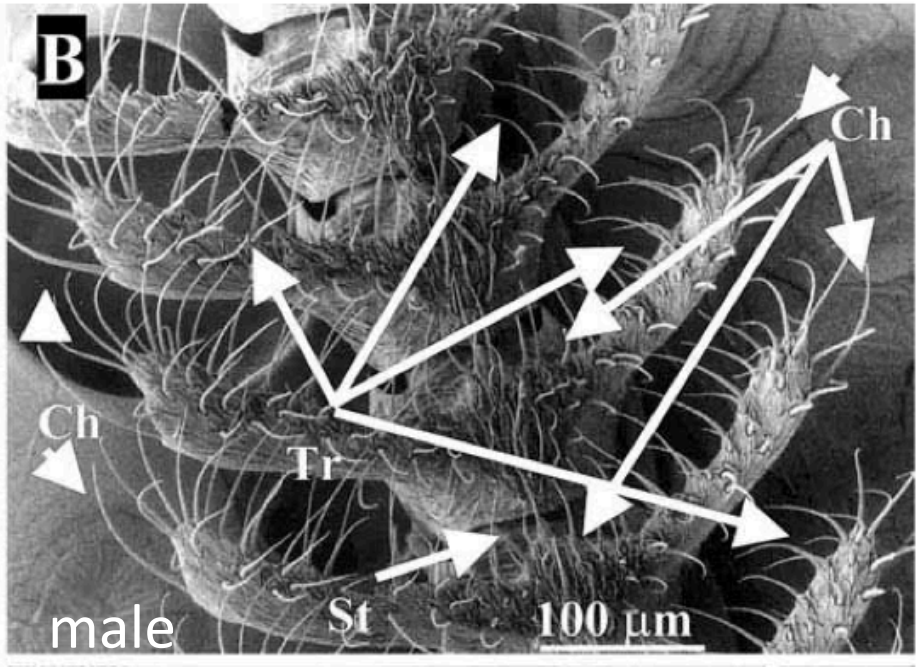


Participants from past bioinformatics and genomics workshops were invited to gain hands-on experience with real data– 54 people from 5 African countries came!

What Have We Found So Far?

Genes upregulated male versus female antennae, endogenous viruses and cases of horizontal transfer

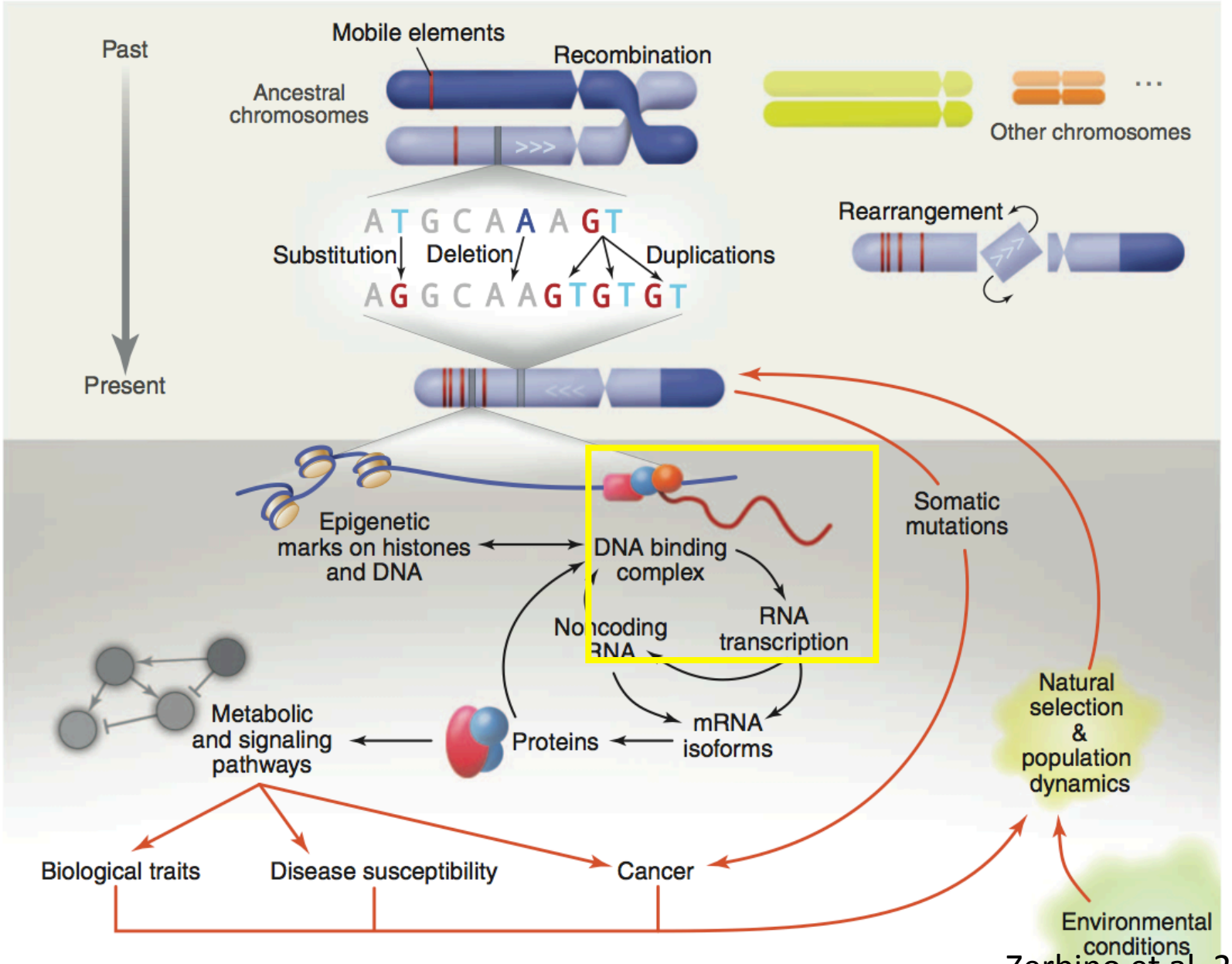




Members of the *Busseola Fusca* Genomics Consortium

Absolomon Kihara
Alan Orth
Alistair Darby
Appolinaire Djikeng
Bruno Leru
Caitlin Miller ('14)
Okalang Uthman
Wacoo Paul
Alexander Ssamula
Muzoora Saphan
Kerfua Susan
Tolbert Sonda
Elizabeth Ajema Luvai
Ernest Lutomia
Chebon Lorna Jemosop
Ngalah Stephen Bidii
Damaris Matoke
Gachara Grace
Harrison Ndung'u
Wachiuri Kelvin
Mwangi
Obange Faith
Towett Sharon
Chepkemoi

Questions in Biology Range from Simple to Complex



Zerbino et al. 2012

Many Studies Focus on Understanding Differences in Gene Expression

Within a Species

Different cells with different jobs

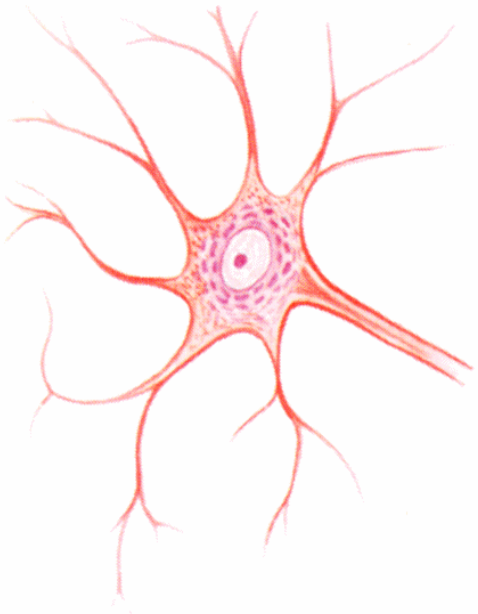
- Neurons
- Skin cells
- Intestinal cells
- Liver cells
- Blood cells
- Fat cells
- Germ cells
- Muscle cells

Same DNA!

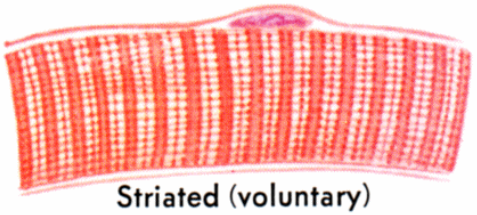
Responses to different conditions

- Heat/Cold
- Food/Hunger
- Disease
- Stress
- Danger
- Development
- Aging
- Damaging chemicals

NERVE CELL

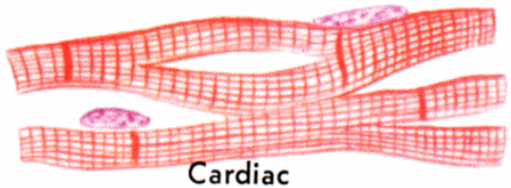


MUSCLE CELLS



Striated (voluntary)

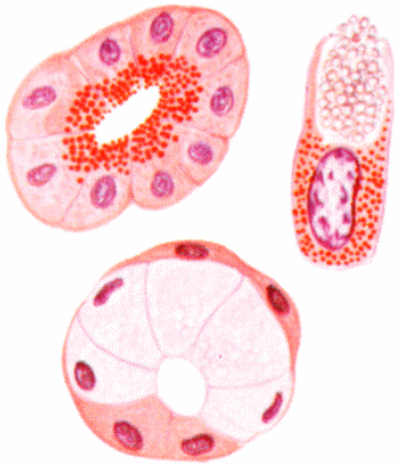
Smooth (involuntary)



Cardiac

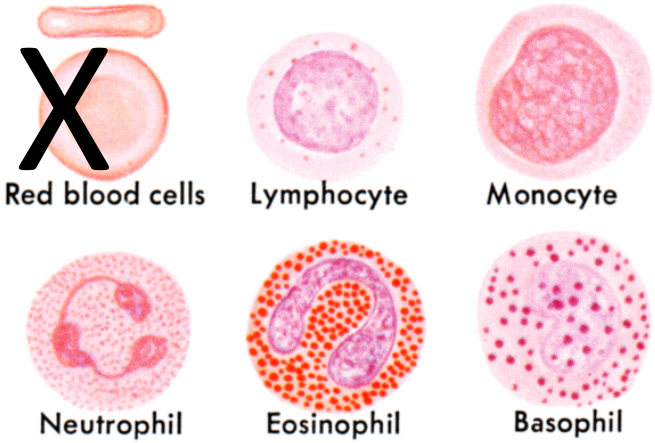
BONE CELL

GLAND CELLS

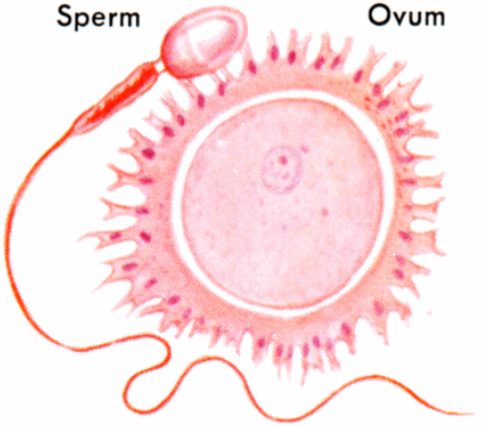


BLOOD CELLS

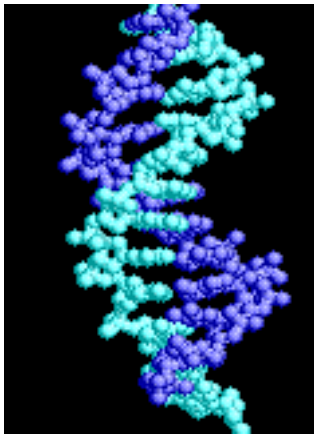
White blood cells

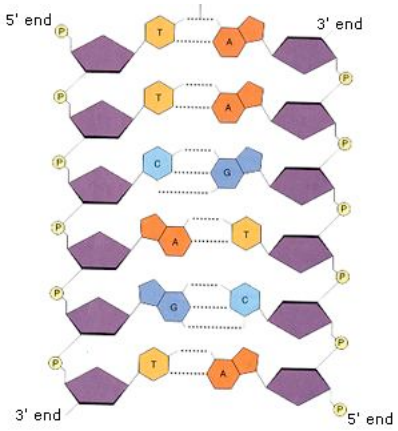


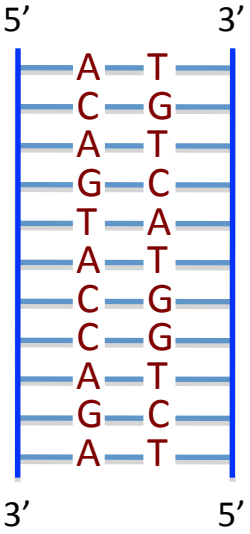
REPRODUCTIVE CELLS



DNA





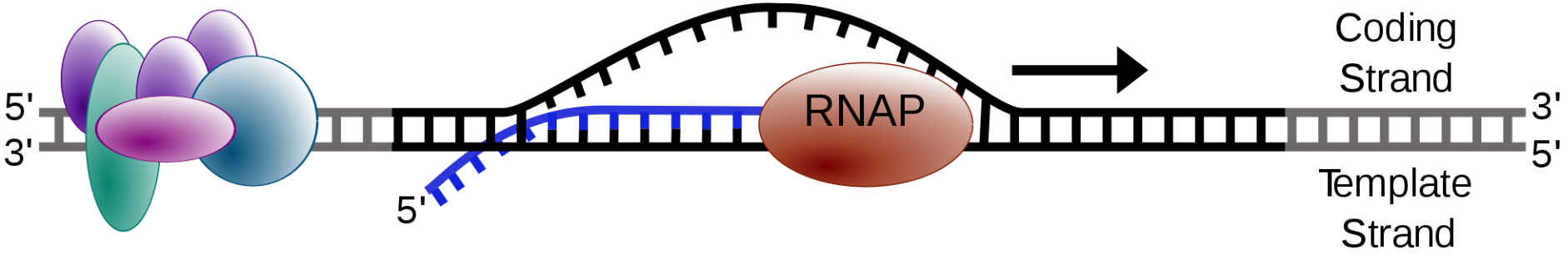


5’ ACAGTACCAGACCAGACCATACATACCATC 3’
3’ TGTCATGGTCTGGTCTGGTATGTATGGTAG 5’

ACAGTACCAGACCAGACCATACATACCATC

RNA

Transcription produces mRNA polymers

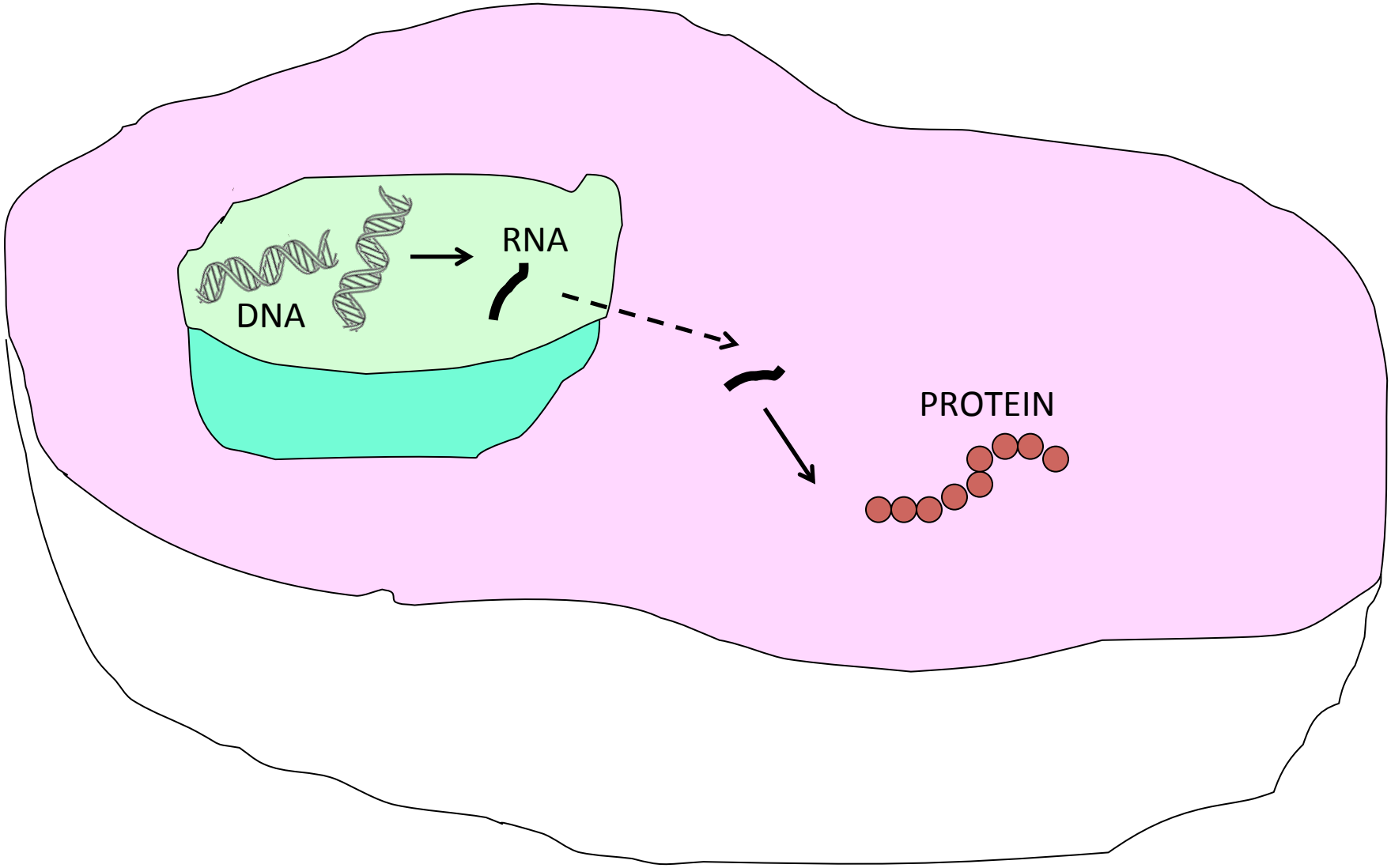


Also rRNAs, tRNA, siRNA, etc.

RNA is typically single-stranded, but can fold back on itself, making small hairpins, and therefore is capable of forming complex structures

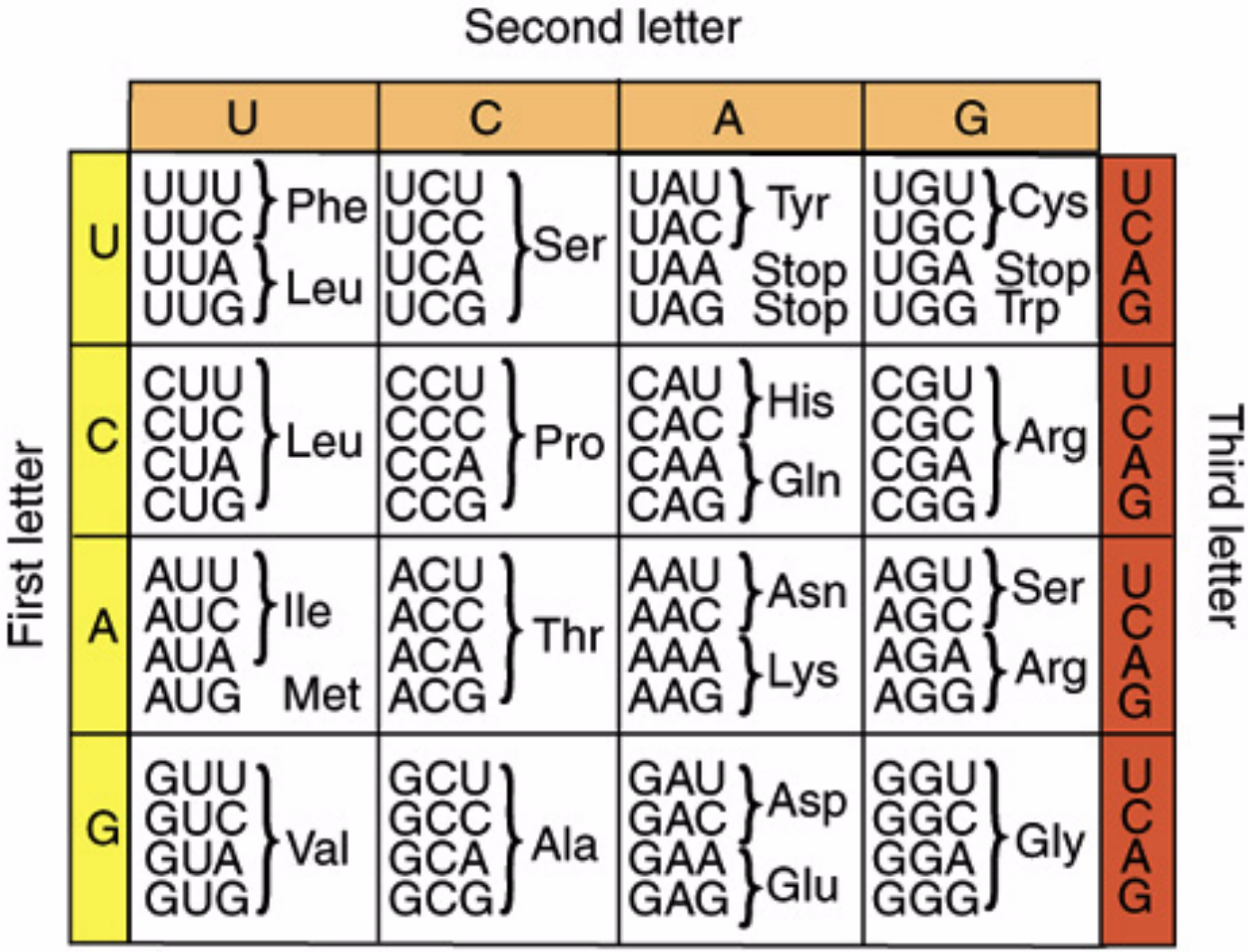
Thus, mRNAs are messengers, but many other RNA molecules are functional and act much like proteins do

THE CENTRAL DOGMA



Coding DNA → proteins → us

The genetic code



Synonymous vs. nonsynonymous differences

Proline
four-fold degenerate amino acid

C C U
C C C
C C A
C C G

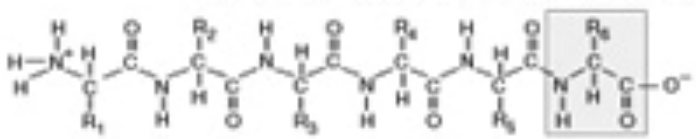
Synonymous changes
Nonsynonmous changes

Arginine

C G T

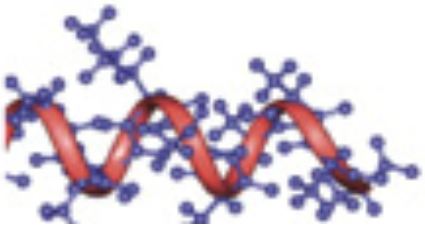
Protein Structure

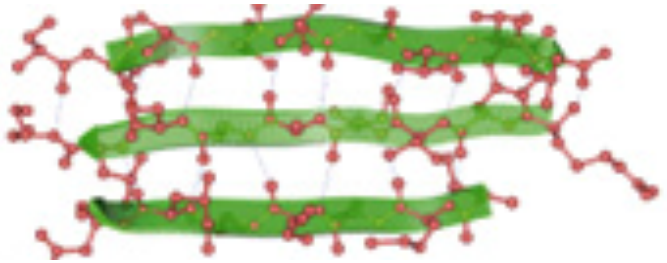
1^0 structure



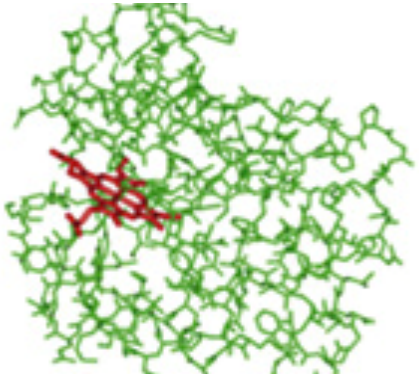
N terminus C terminus

2^0 structure

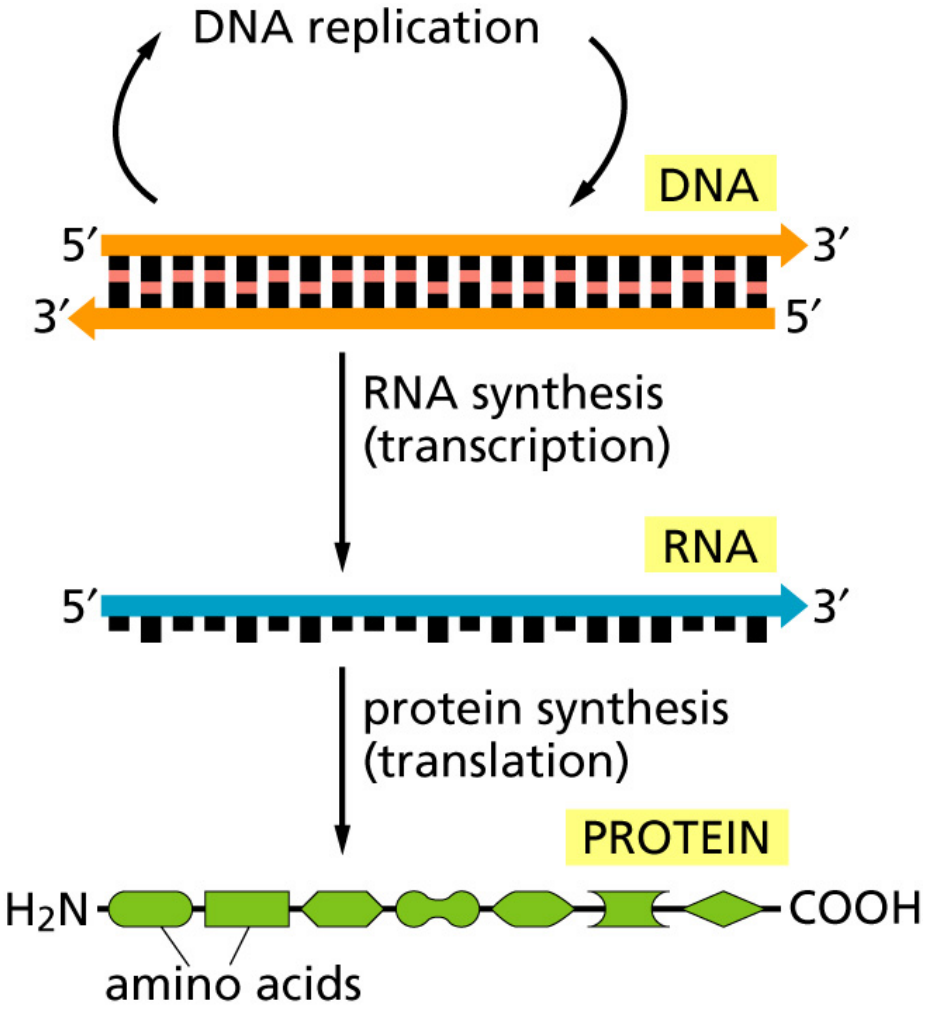




3^0 structure

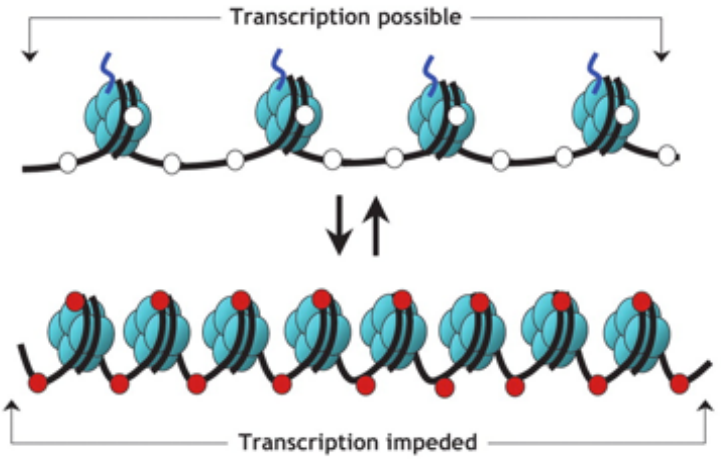


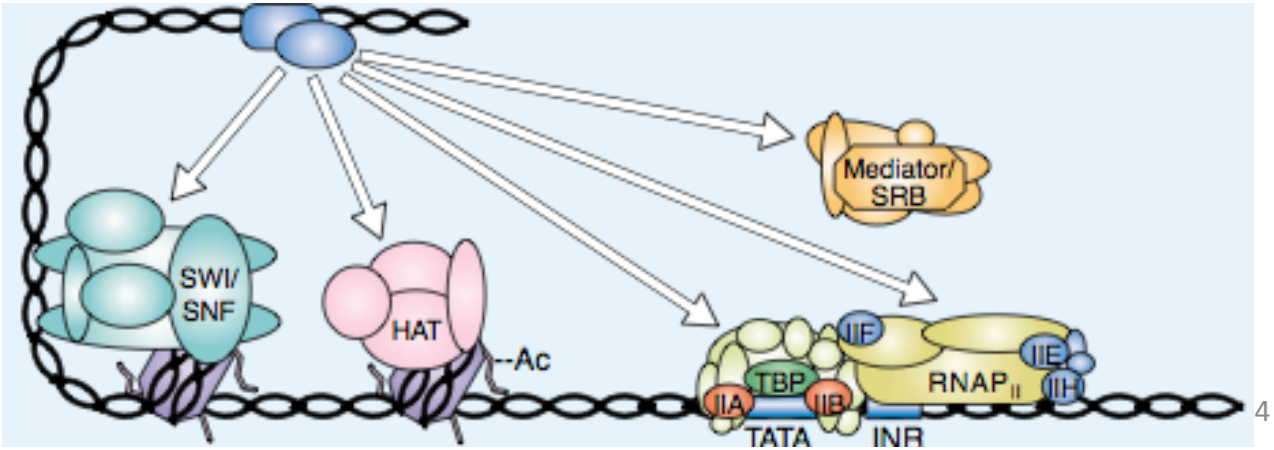
Central Dogma

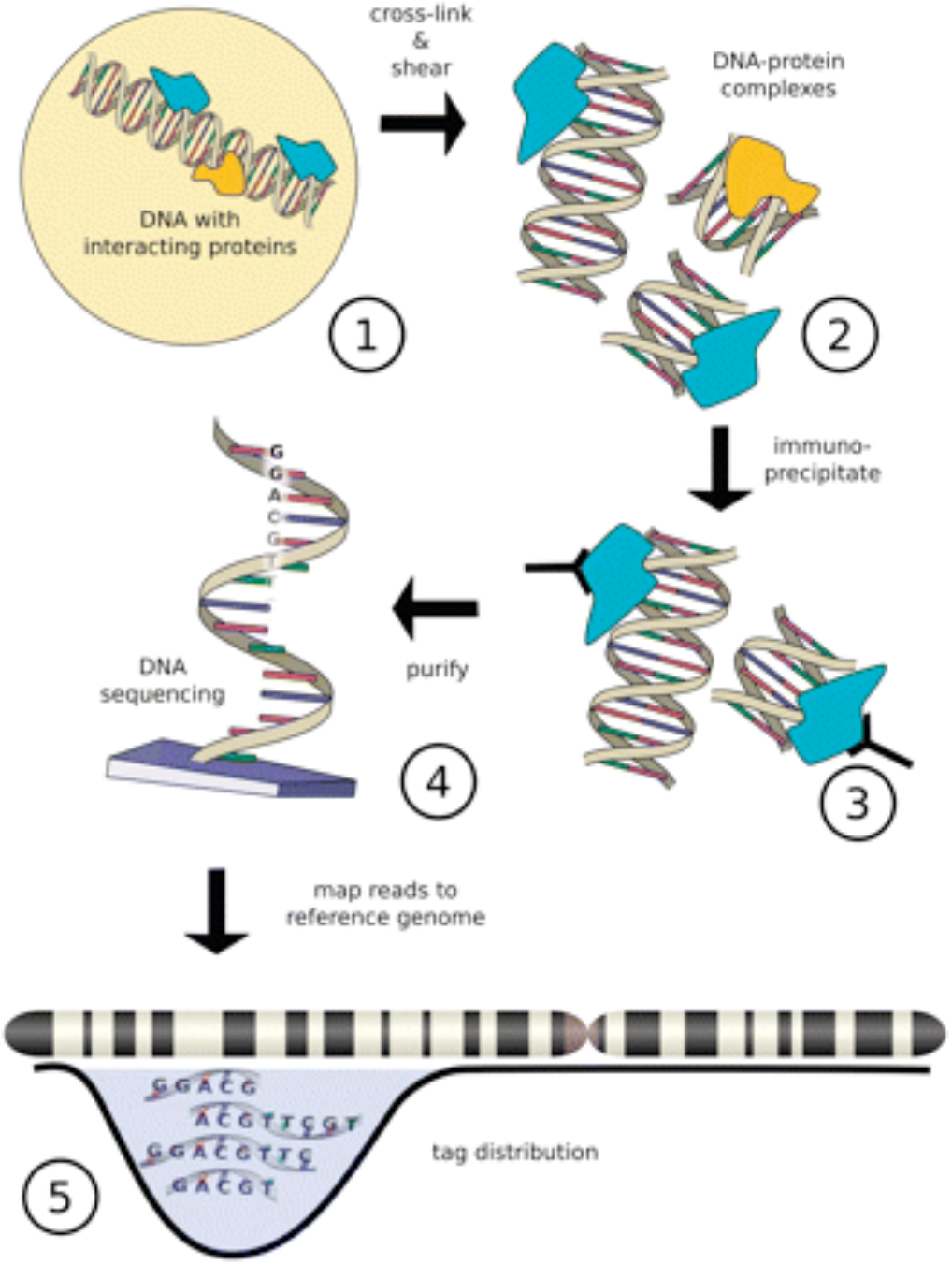


How Can Transcription Levels Be Quantified Using Bioinformatics and Genomics?

- DNA availability (methylation and acetylation)
  - Bisulfite sequencing
- Transcription Initiation and Elongation
  - ChIP-Seq
- Transcript termination, processing and export
  - RNA-Seq







ChIP-Seq

- Take cells of interest and cross-link (fix proteins to the DNA that they bind to while "in action")
- Shear DNA
- Immunoprecipitate YPOI (your protein of interest, here the blue protein) and discard everything else
- Release the protein from the fragments it was bound to
- Sequence the DNA
- Map the data to the genome to see what genes nearby might be regulated by YPOI

LINK TO ORIGINAL ARTICLE

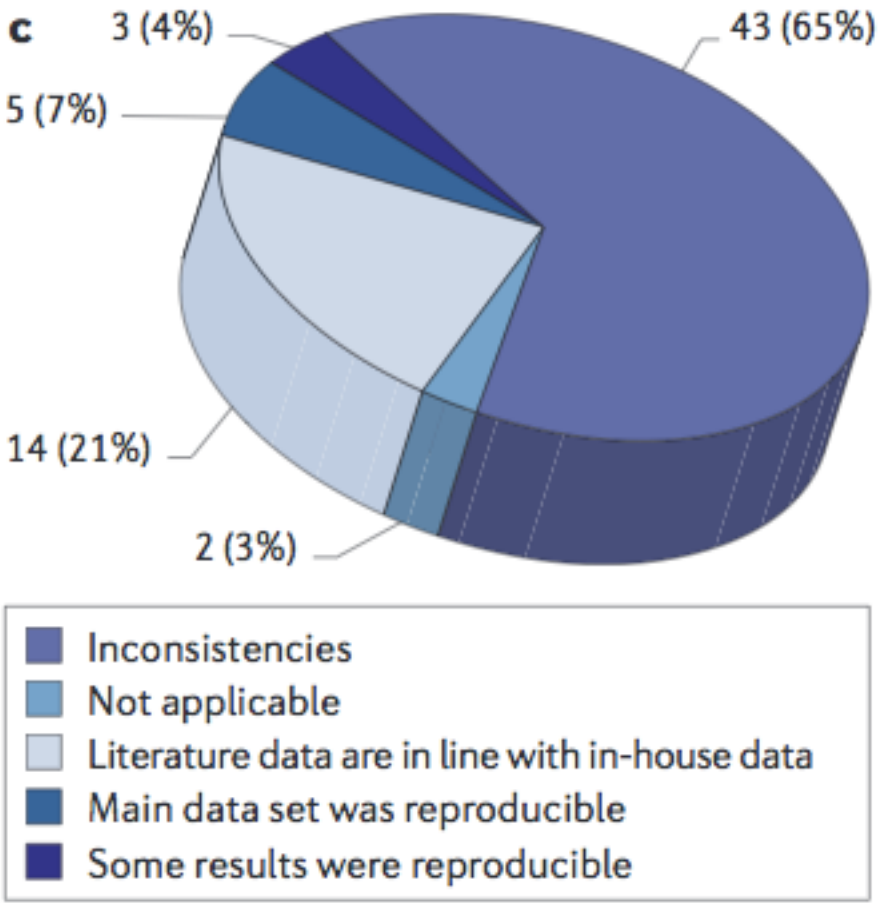
Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

A recent report by Arrowsmith noted that the success rates for new development projects in Phase II trials have fallen from 28% to 18% in recent years, with insufficient efficacy being the most frequent reason for failure (Phase II failures: 2008–2010. *Nature Rev. Drug Discov.* **10**, 328–329 (2011))[1]. This indicates the limi-

to 'feasible/mar
of pursuing a f
development p
get could ultim
Euros. Even in
in activities su
ing programm

results that are published are hard to reproduce. However, there is an imbalance between this apparently widespread impression and its public recognition (for example, see REFS 2,3), and the surprisingly few scientific publications dealing with this topic. Indeed, to our knowledge, so far there has been no published



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, **2** Centre for Cancer Biomedicine, University of Oslo, Blindern, Oslo, Norway, **3** Department of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University Park, Pennsylvania, United States of America, **4** Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America, **5** Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway, **6** Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

Replication is the cornerstone of a cumulative science [1]. However, new

We further note that reproducibility is just as much about the habits that ensure

than to do it while underway). We believe that the rewards of reproducibility will

10 Simple Rules for Reproducibility

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results