UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Research Parasite & Data Re-Use: Bacteria-Animal Lateral Gene Transfer

Julie C. Dunning Hotopp
Institute for Genome Science
Department of Microbiology and Immunology

Research Parasites

- Researchers who generate new hypotheses from existing data
- 2016 New England Journal of Medicine editorial termed these researchers as "research parasites”
  - tongue-in-cheek name
- The Parasite awards, given annually, recognize outstanding contributions to the rigorous secondary analysis of data.

Companion Award: Research Symbiont Awards for data generators that encourage open data

Most Famous Case of Lateral Gene Transfer (LGT)

Real Advantageous Eukaryote LGT

- Thought to be uncommon
- Less frequent in animals

Color polymorphism by carotenoid biosynthesis acquired from fungi

Coffee berry parasitism by beetle acquiring *Bacillus* mannanase

Moran & Jarvik, Science, 2010; Acuna et al., PNAS, 2011.

Brown Marmorated Stinkbugs

- *Halyomorpha halys*
  - Hemiptera: Pentatomidae
- Invasive pest
- Native to Asia
- First observed in Allentown, PA in 1996
- Multiple bacterial mannanases

Ioannidis et al., BMC Genomics, 2014.

More Real Advantageous Eukaryote LGT

- Cellulases and other plant cell wall degrading enzymes in plant parasitic nematodes

e.g. Danchin et al., PNAS, 2010.

Serial Endosymbiosis Theory (SET)

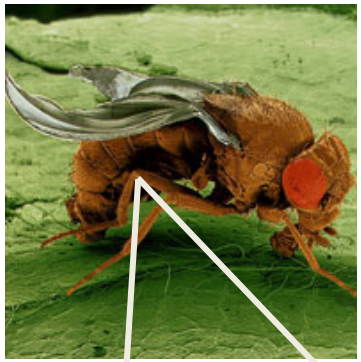
- Explains the acquisition of mitochondria and chloroplasts by eukaryotes
- Over time, the accumulation of endosymbiont genes in the nuclear genome, combined with organelle protein uptake systems, enable the transition of an endosymbiont to an organelle.

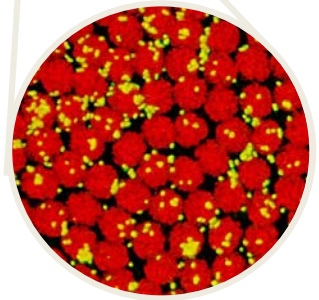
Margulis, L. 1970. Origin of Eukaryotic Cells.

The Many Hosts of Wolbachia Endosymbionts

Research Parasitism #1 (ca. 2005) -- Serendipitous Genome Discovery

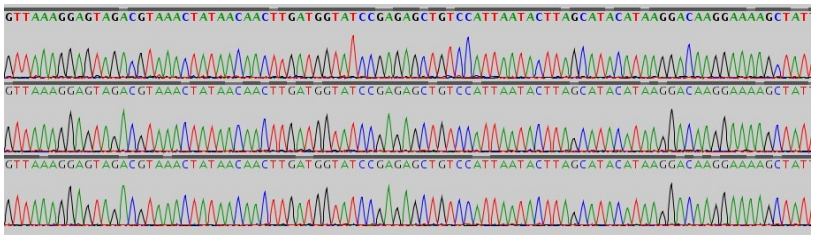
Wolbachia-infected *Drosophila*



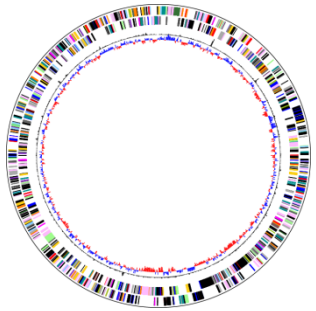


library construction

sequencing



assembly



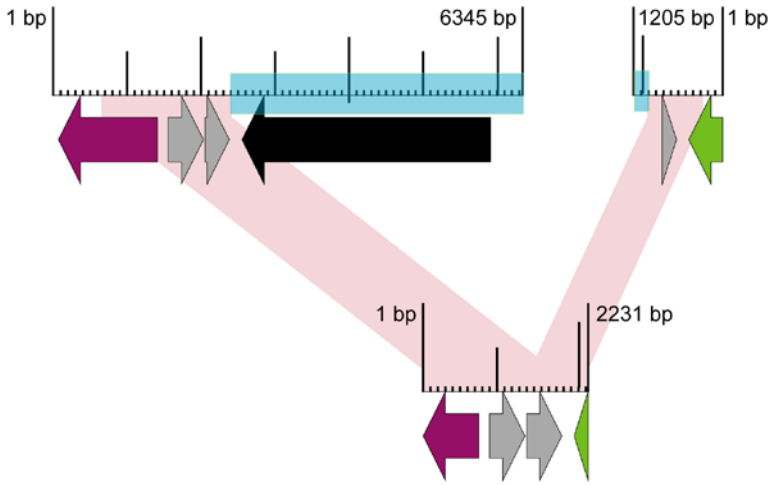
Wolbachia

Drosophila

Salzberg et al. (2005) Genome Biology 6(3):R23.

wAna Anomalies

- Find *Drosophila* ORFs in the *Wolbachia* genome
- Multiple copies of the same gene
  - one intact
  - one disrupted by *Drosophila* retroelement



wAna Anomalies

Possibilities:

H1: Chimeric libraries were sequenced

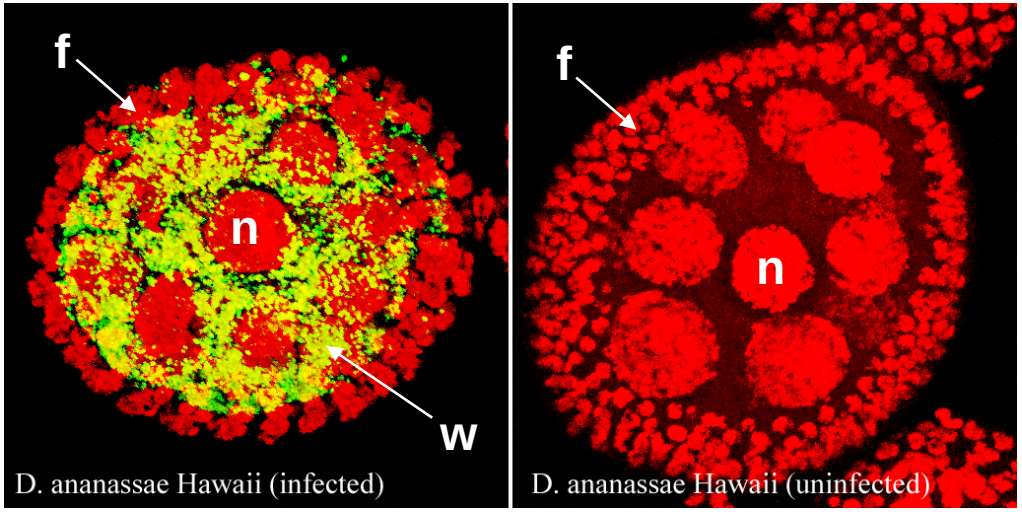
H2: Interdomain lateral gene transfer

In this case, we can just order the flies and perform follow-up experiments.

Extensive transfer of the genome

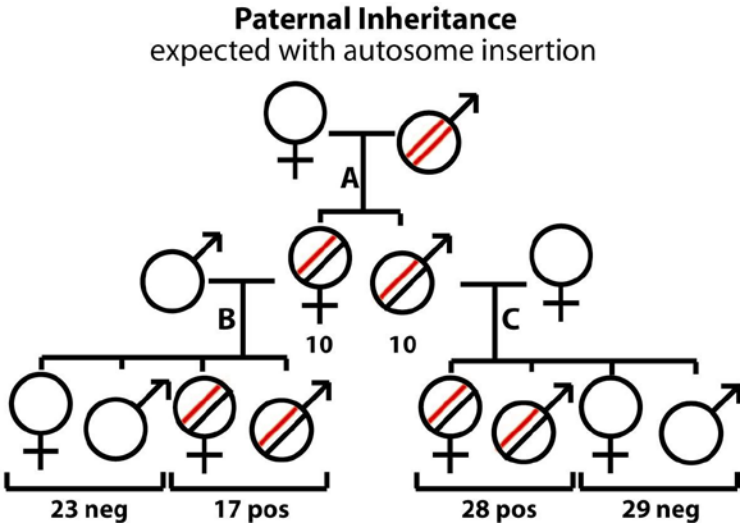
	Infected	Uninfected
Hawaii (insert)	43/46	43/46
Townsville (no insert)	Not tested	1/46

f = follicle cell nuclei
n = nurse cell nuclei
w = *Wolbachia*



Infected **Uninfected**

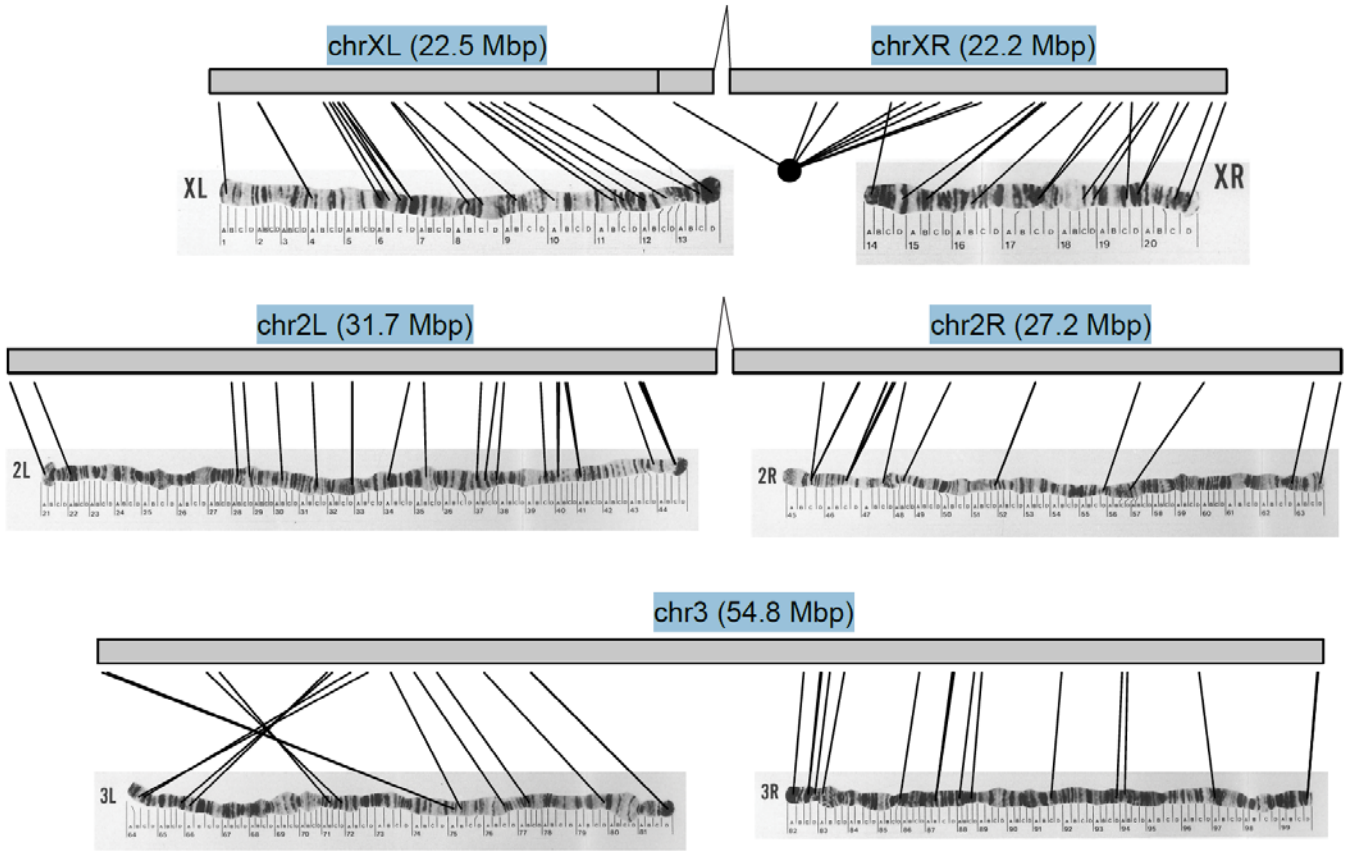
Inheritance on a single autsome



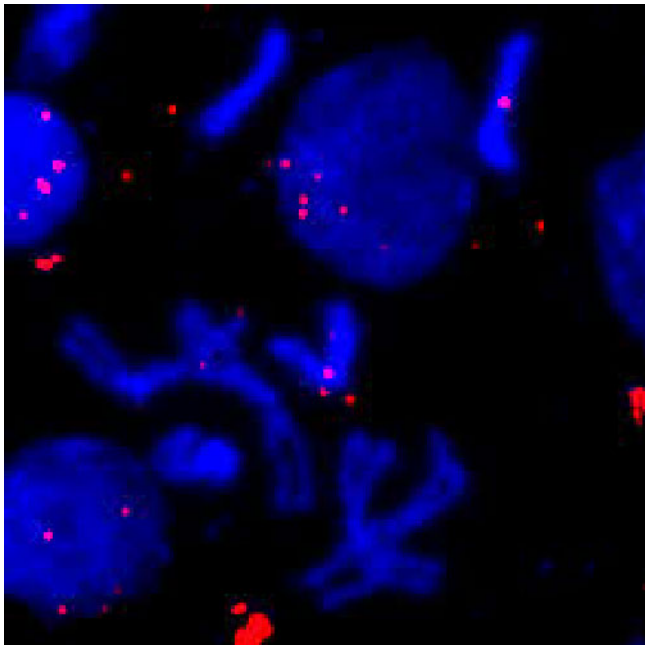
28/57 (49%) of offspring are positive
15/28 (54%) males
13/29 (45%) females

Three loci show segregation: 16S rRNA, *wsp*, *gatB*

New Nearly Complete Genome using PacBio Sequel2 and MinION RAD



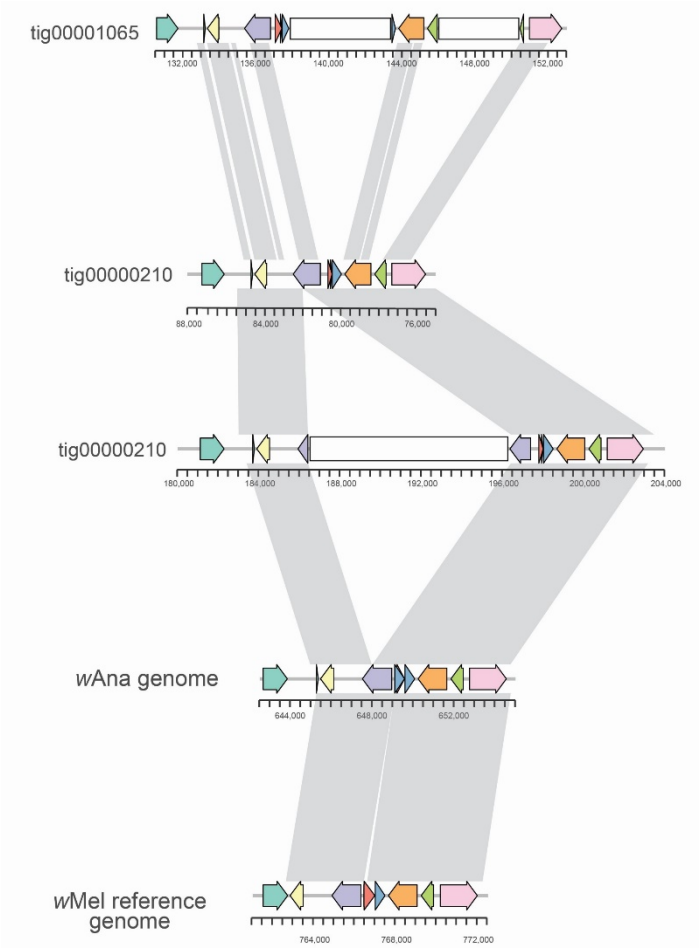
FISH with Nuwts



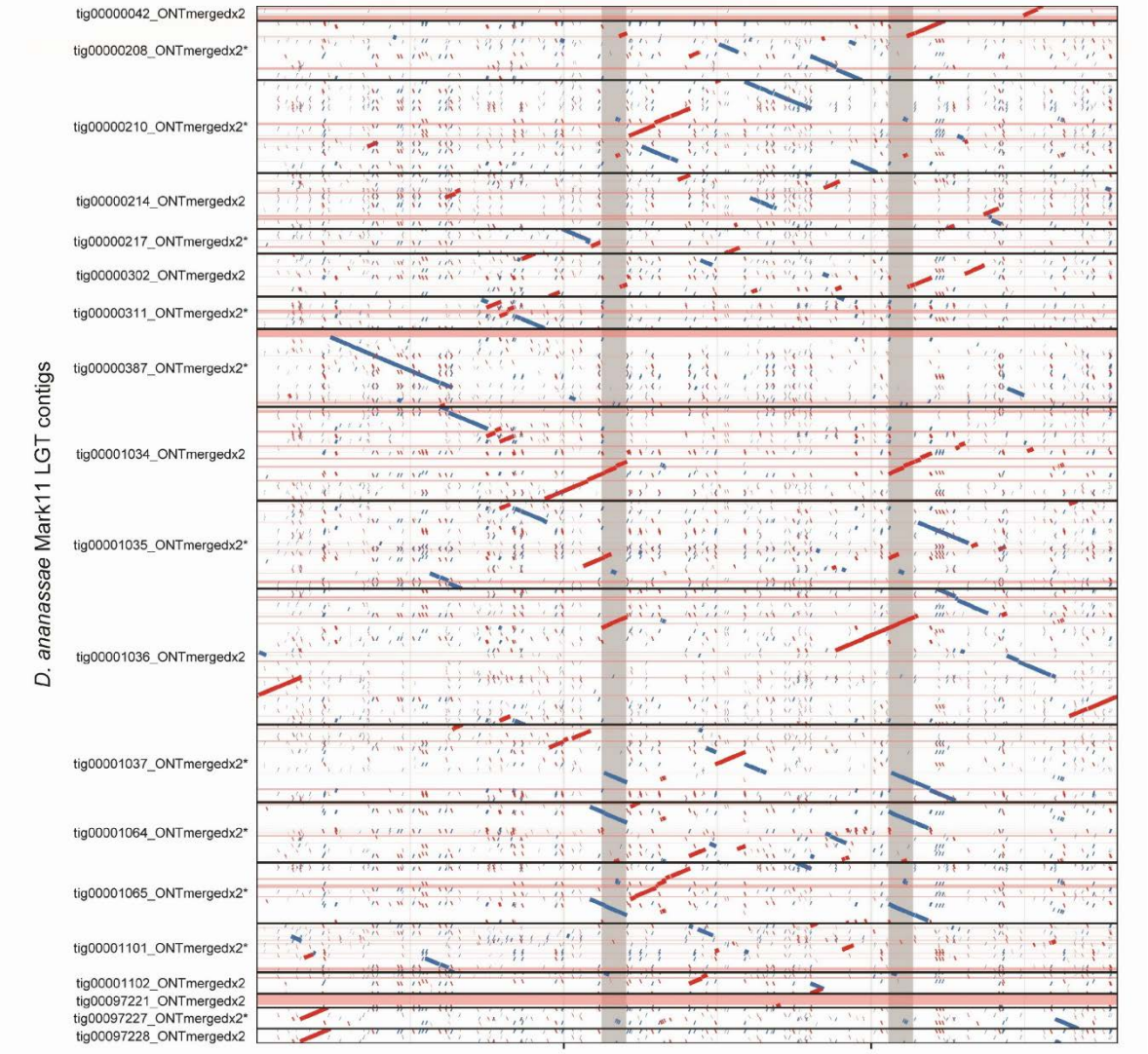
- LGT is in the abnormally large and heterochromatic 4th Chromosome
- Multiple sites hybridize
- >2% of the *D. ananassae* genome is derived from *Wolbachia* DNA
- 20% of chromosome 4 in *D. ananassae* is derived from *Wolbachia* DNA

Klasson et al., BMC Genomics, 2014.

Three Copies of a Nuwt—Different Outcomes



Very Fragmented, Because Massive Repeat



G3: GENES, GENOMES, GENETICS *August 1, 2017 vol. 7 no. 8 2439-2460*

G3 Genes | Genomes | Genetics

‹ Previous Article Next Article ›

Our estimate is ~1000X higher.

Retrotransposons Are the Major Contributors to the Expansion of the *Drosophila ananassae* Muller F Element

Wilson Leung, Christopher D. Shaffer, Elizabet
John M. Braverman, Thomas C. Giarla, Nathar
Srebrenka Robic, Shannon R. McCartha, Danie

Abstract

The discordance between genome size and the complexity of eukaryotes can partly be attributed to differences in repeat density. The Muller F element (~5.2 Mb) is the smallest chromosome in *Drosophila melanogaster*, but it is substantially larger (>18.7 Mb) in *D. ananassae*. To identify the major contributors to the expansion of the F element and to assess their impact, we improved the genome sequence and annotated the genes in a 1.4-Mb region of the *D. ananassae* F element, and a 1.7-Mb region from the D element for comparison. We find that transposons (particularly LTR and LINE retrotransposons) are major contributors to this expansion (78.6%), while *Wolbachia* sequences integrated into the *D. ananassae* genome are minor contributors (0.02%). Both *D. melanogaster* and *D. ananassae* F-element genes exhibit distinct characteristics

Problems with Data Re-use

- Lack of adequate reporting of methods
  - Data cleansing
  - Contamination removal
  - Disappearance of collapsed repeats (e.g in degens)
  - Over-emphasis on the reliability of a consensus genome

Research Parasitism #2 (ca. 2006) – Scanning the trace repositories

- 26 arthropod and filarial nematode genomes
  - Have potential of being *Wolbachia*-infected
- 15 are organisms known to be infected
  - 20-70% of arthropods in the wild are infected
- 10 of these organisms have *Wolbachia* traces
- 8 show evidence of *Wolbachia*-host LGT

Dunning Hotopp et al. (2007) Science 317: 1753-6.

Wolbachia-host LGT Prevalence

- 31% of potentially infected organisms have LGT

AND

- 80% of genomes with *Wolbachia* reads have LGT

Caveats:

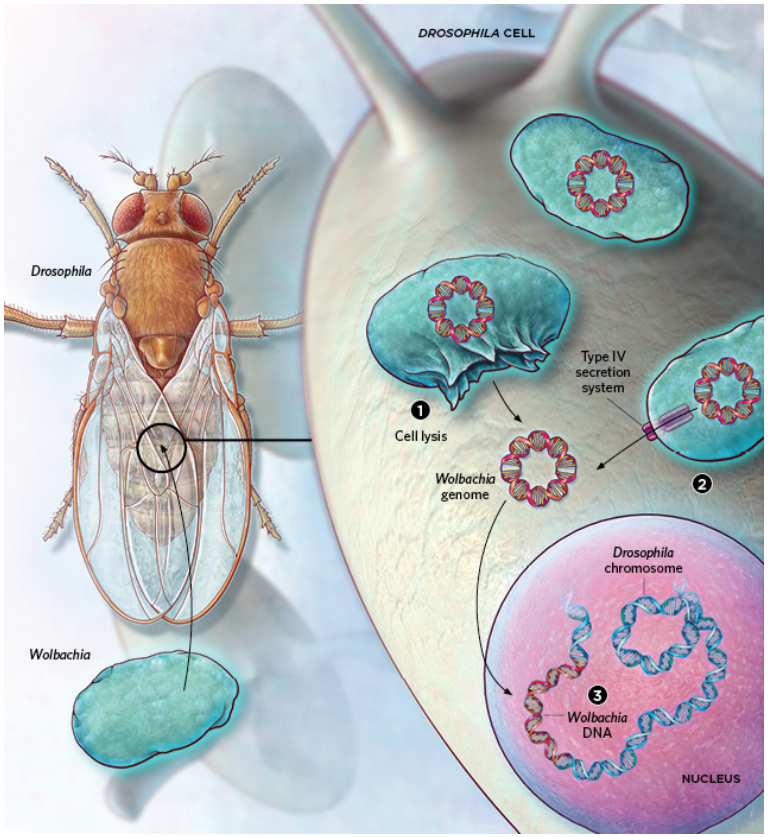
Not a random sampling of organisms

Deposited traces may be cleansed of bacterial traces

Not all genomes are deposited

Nuwts: Nuclear *Wolbachia* Transfers

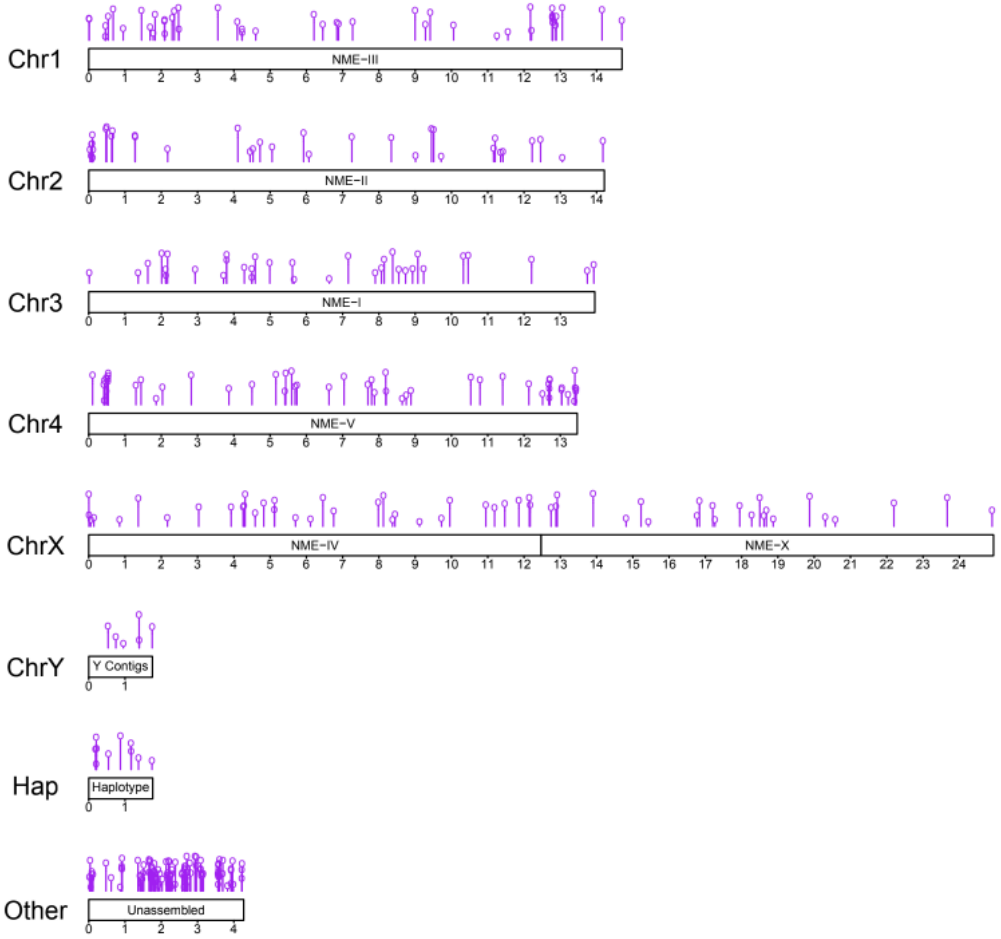
- 1/3 insect/nematode genomes sequenced in 2007 had nuwts
- Most endosymbionts do NOT do this
  - Stem-cell associated endosymbionts?
- In some insects, the entire *Wolbachia* genome integrated



Robinson and Dunning Hotopp. The Scientist, 2016.

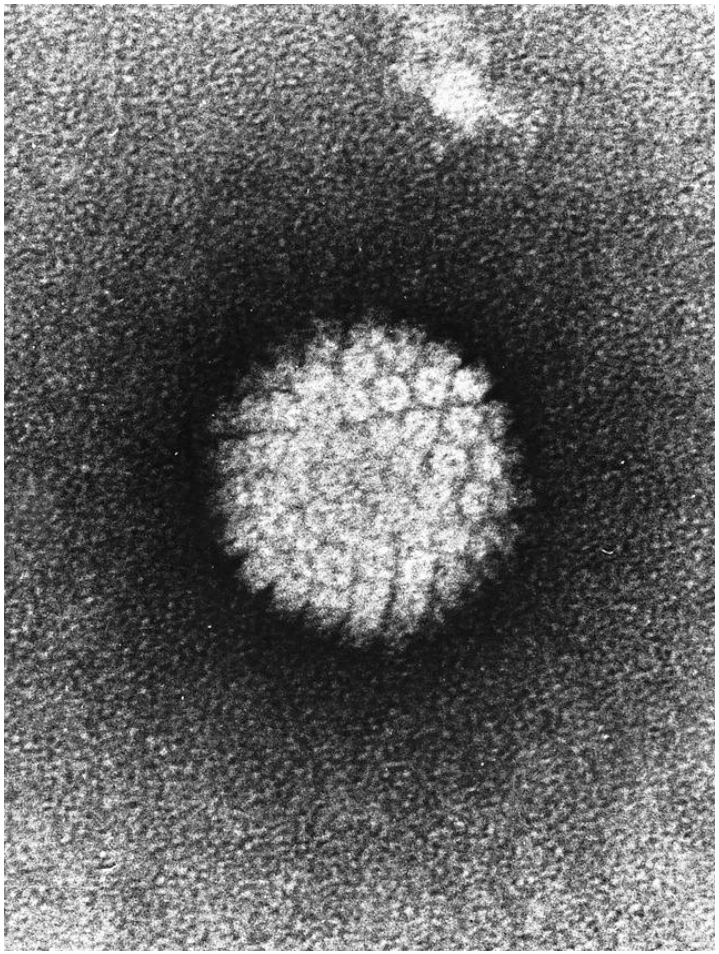
*w*Bm LGT into *B. malayi*

- 345 *B. malayi* regions
  - Spans ~428 kbp
  - Largest: 29.7 kbp
  - 0.4% of genome
- 133 *w*Bm regions
  - Spans ~144 kbp
  - ~1.4% of the *Wolbachia* genome
- Thus, most are multicopy "repeats"
  - 59 are present >1
  - 5 have >10 copies
- Most are frameshifted
  - Only 21 full-length protein-coding regions, and many of those have altered start and stop codons
- Frequently in unscaffolded contigs



LGT can be beneficial and potentially neutral. Can they be deleterious?

Deleterious LGT—HPV



Crown gall disease in plants –*Agrobacterium tumefaciens*

- Directed transfer
  - 10-30 kbp of T-DNA from its Ti plasmid (200-800 kbp) to plants
- Type IV secretion system
- Targeted to the nucleus
- Incorporated by illegitimate recombination
- Transcription from T-DNA encoded eukaryotic promoters

Question

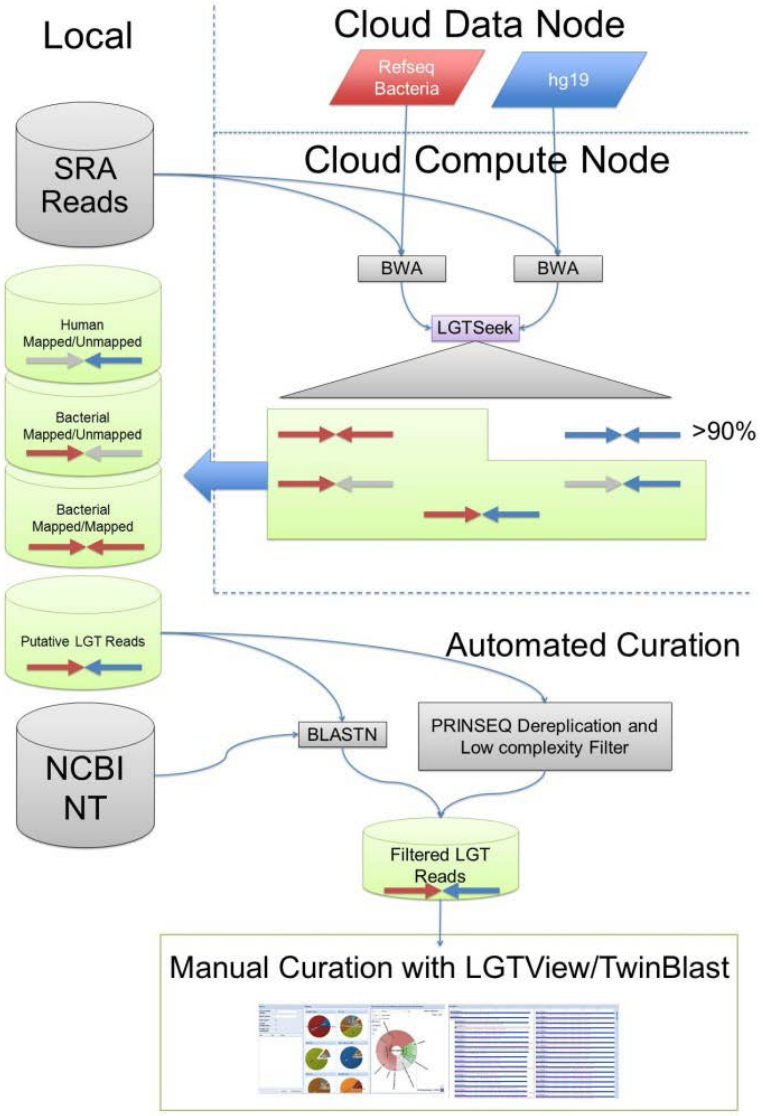
- If LGT is so prevalent from bacteria to invertebrates, is it also prevalent in other animals, like mammals?
- If so, is the lack of inheritance of LGT merely due to a lack of LGT in germ cells?
  - This may be the case in invertebrates as well since a germ line endosymbiont is what participates in this phenomenon widely
- Can transfers happen frequently in somatic cells where they would mutagenize the genome?

Microbial infections and cancer

- There are 10x more bacterial cells in our bodies than human cells.
- Worldwide, 15-20% of cancers are linked to bacterial, viral, or parasitic infections.

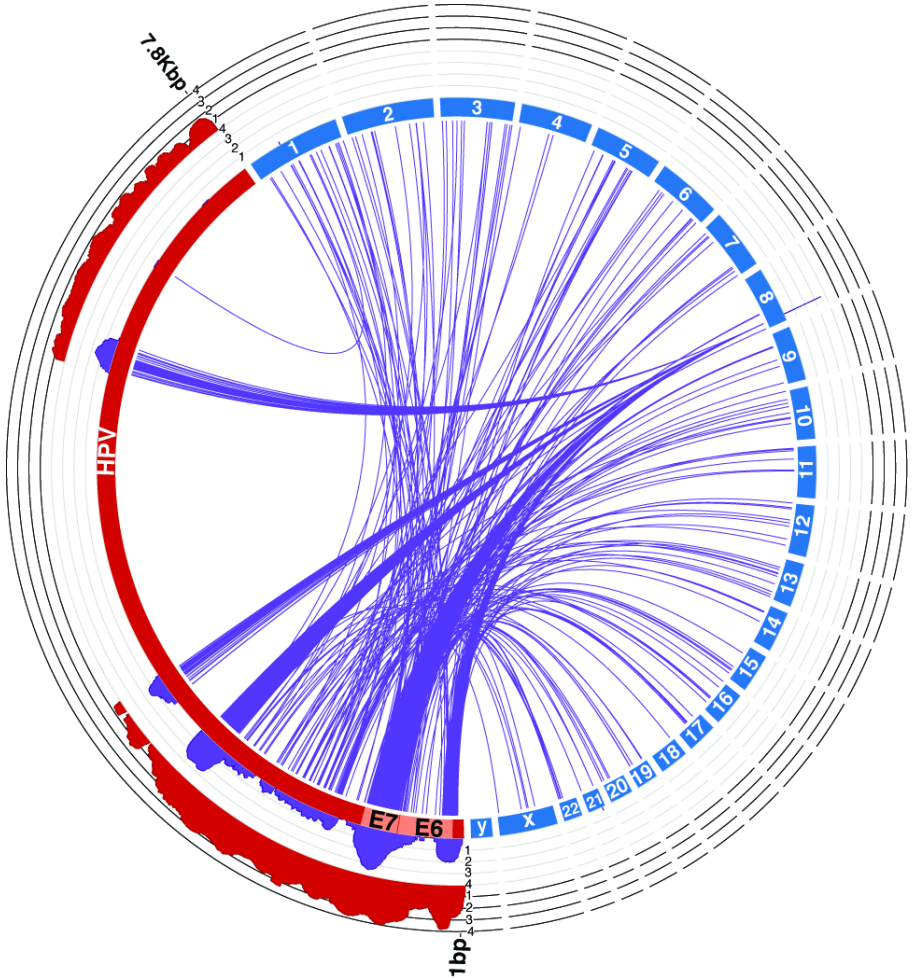
Research Parasitism #3 (ca. 2013) – Scanning TCGA data

- Nobody was going to fund me to sequence cancer samples in the hopes of finding deleterious LGT.
- The Cancer Genome Atlas (TCGA) was sequencing 1000+ cancer samples and providing the data to the public



Riley, Sieber, PLoS Comp. Bio., 2013.

Calibration with HPV in HeLa



- 6,333 reads supporting integration of HPV into the human genome
  - 0.12% of the total read pairs
  - Flank the constitutively expressed E6 and E7 viral oncogenes.
  - Vast majority comes from the known tandem integration site on chromosome 8

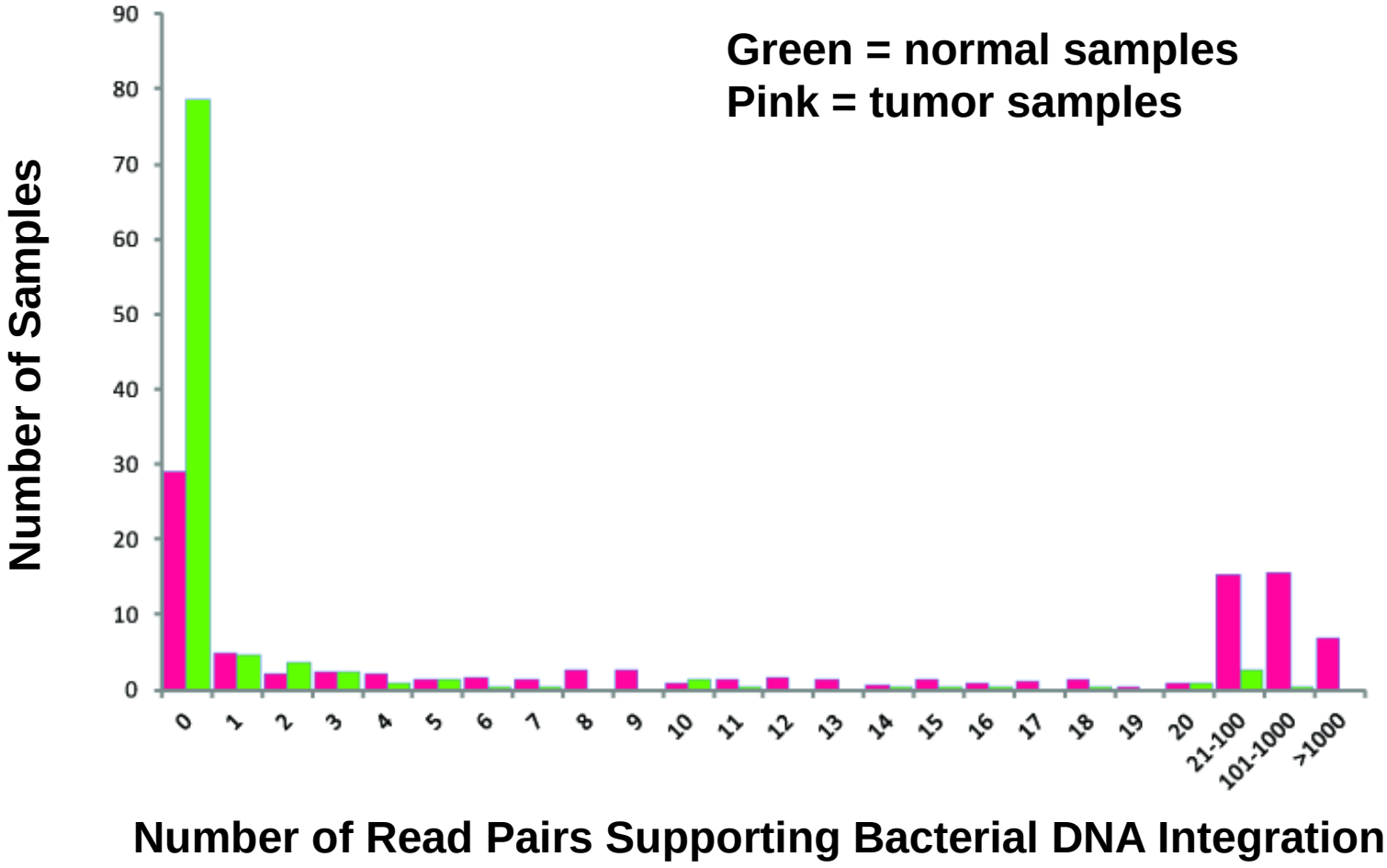
Riley, Sieber, PLoS Comp. Bio., 2013.

The Cancer Genome Atlas (TCGA)

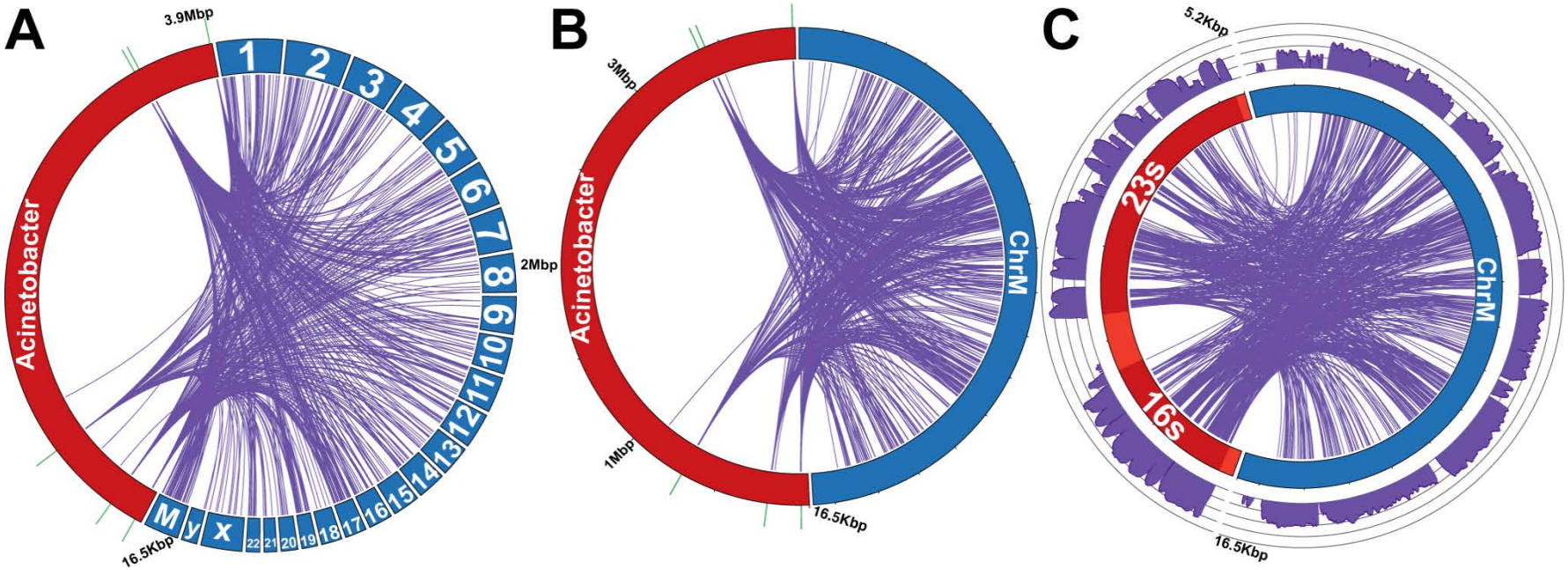
- 6.6 trillion bases of Illumina paired-end sequencing data
- 691,560 read pairs supporting bacterial integration
  - 1× to 150× coverage.
- 63.5% of the TCGA analyzed were tumor samples
  - 99.9% of reads supporting bacterial integration came from tumor samples
- Majority of normal samples had no read pairs supporting integrations
  - Majority of tumor samples had >10 reads supporting integrations

Riley, Sieber, PLoS Comp. Bio., 2013.

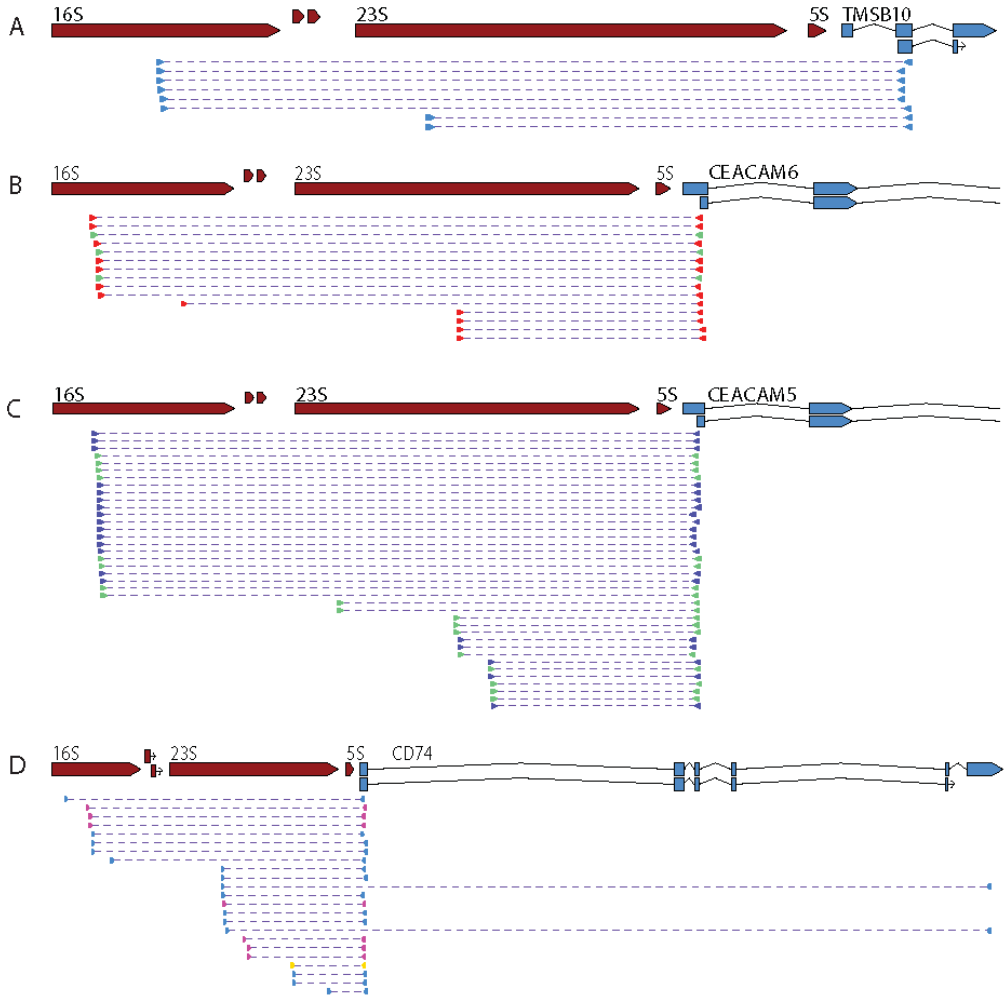
Distribution in tumor v. normal



Acute Myeloid Leukemia (LAML)



Stomach Adenocarcinomas (STAD)



Riley, Sieber, PLoS Comp. Bio., 2013.

Problems Along the way

- Solely computational research
- Experimental validation needed, but no access to samples
- Problems with the data accessibility statements
  - Statements said one thing
  - Interpreted differently or differentially

Summary

- **Pros:**
  - Data reuse, a.k.a. research parasitism, is a great way to test new and controversial hypotheses
  - Great for providing preliminary data for a grant
- **Cons:**
  - Frequent problems with understanding methods applied and how they may alter the interpretation
    - Discrepancy between database and manuscript methods
  - Access to samples for validation and follow-up experiments are frequently limited or impossible
  - Sometimes people throw away the data you are interested in, to save space or “clean-up” the data
    - Data sometimes is stored in a manner that is great for existing ideas and hypotheses (e.g. bam files if only mapped reads for SNP-based analyses) but that eliminates the testing of new paradigms

Funding

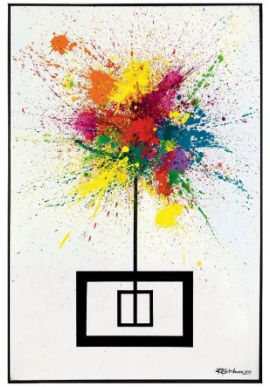
NATIONAL CANCER INSTITUTE

NIH DIRECTOR'S NEW INNOVATOR AWARD

NIAID

U19 AI110820

NIH DIRECTOR'S



TRANSFORMATIVE RESEARCH AWARD

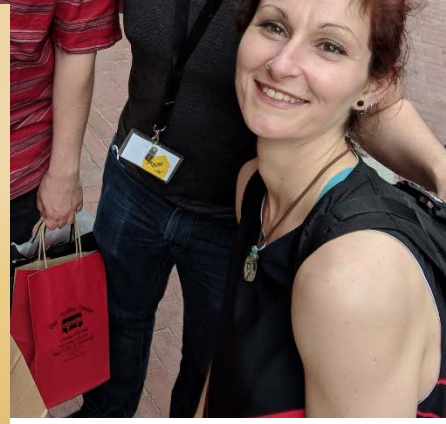
UNIVERSITY of MARYLAND SCHOOL OF MEDICINE

NIH National Institute of Diabetes and Digestive and Kidney Diseases

USDA

BILL & MELINDA GATES foundation

NSF National Science Foundation WHERE DISCOVERIES BEGIN



J D H
L A B
I G S
T

If you are interested in sharing information on lateral gen transfer: we have an NSF-supported YouTube channel on LGT

https://goo.gl/i967FA

Or subscribe: “JDH Lab” on YouTube