Dovetail GENOMICS

# Heaps of Chromosomes

## New Scales and Evolving Paradigms in Genome Assembly

# The Who and the What

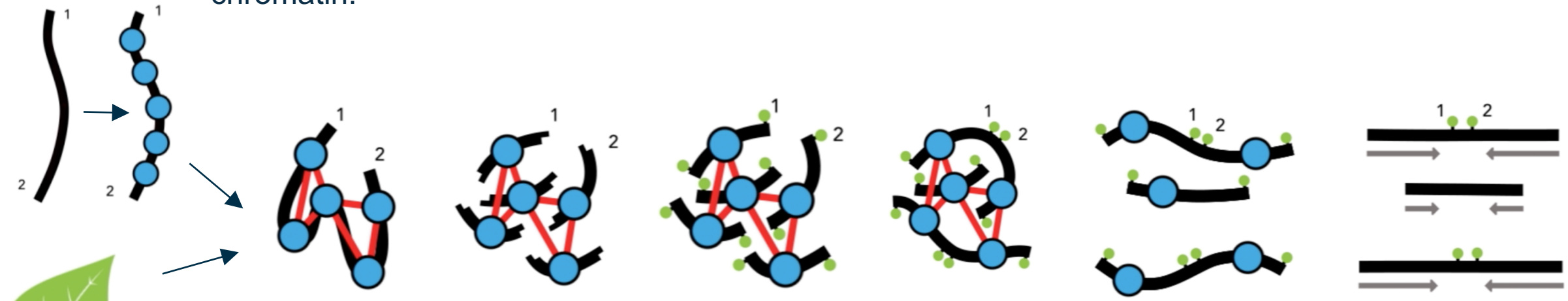
- Genome assemblers
  - Library prep and sequencing
  - Full *de novo* assembly
  - Scaffolding
- Fully integrated, from sample to chromosomes
- Proximity ligation specialists
- Over 200 genomes assembled

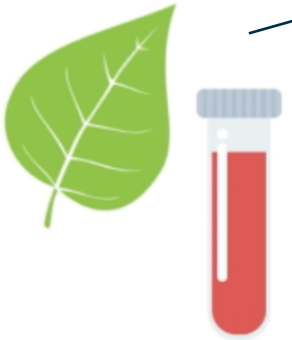
# The (Expanded) Menu

- HMW DNA extraction
- Library prep and sequencing
  - Illumina shotgun
  - **PacBio**
  - Chicago
  - **Dovetail Hi-C**
- *De novo* assembly
  - Meraculous (Illumina)
  - **Falcon (PacBio)**
- Scaffolding
  - HiRise for any proximity ligation data type (Chicago or **Dovetail Hi-C**)
- Gap filling

# Proximity Ligation Approaches

**Chicago™** libraries start from pure DNA that is reconstituted into chromatin.

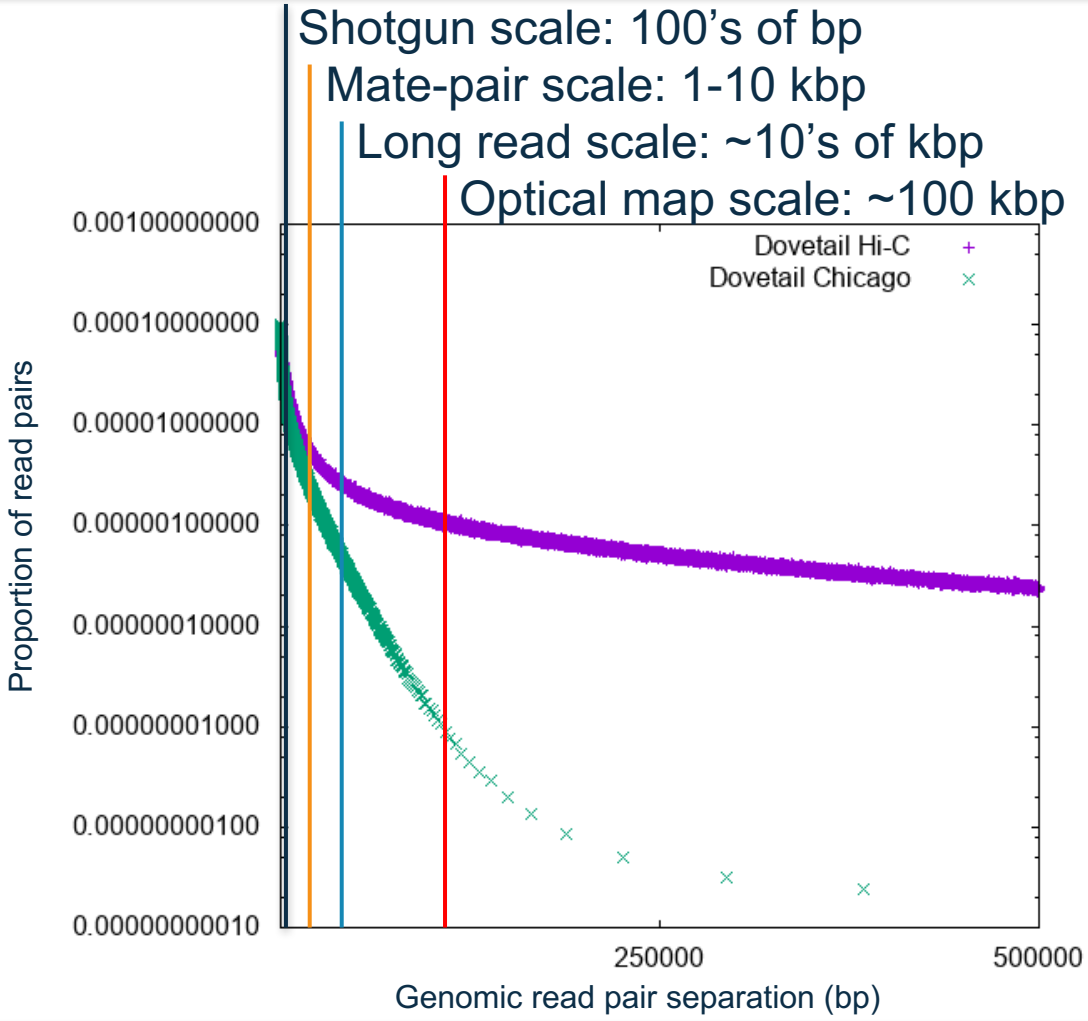


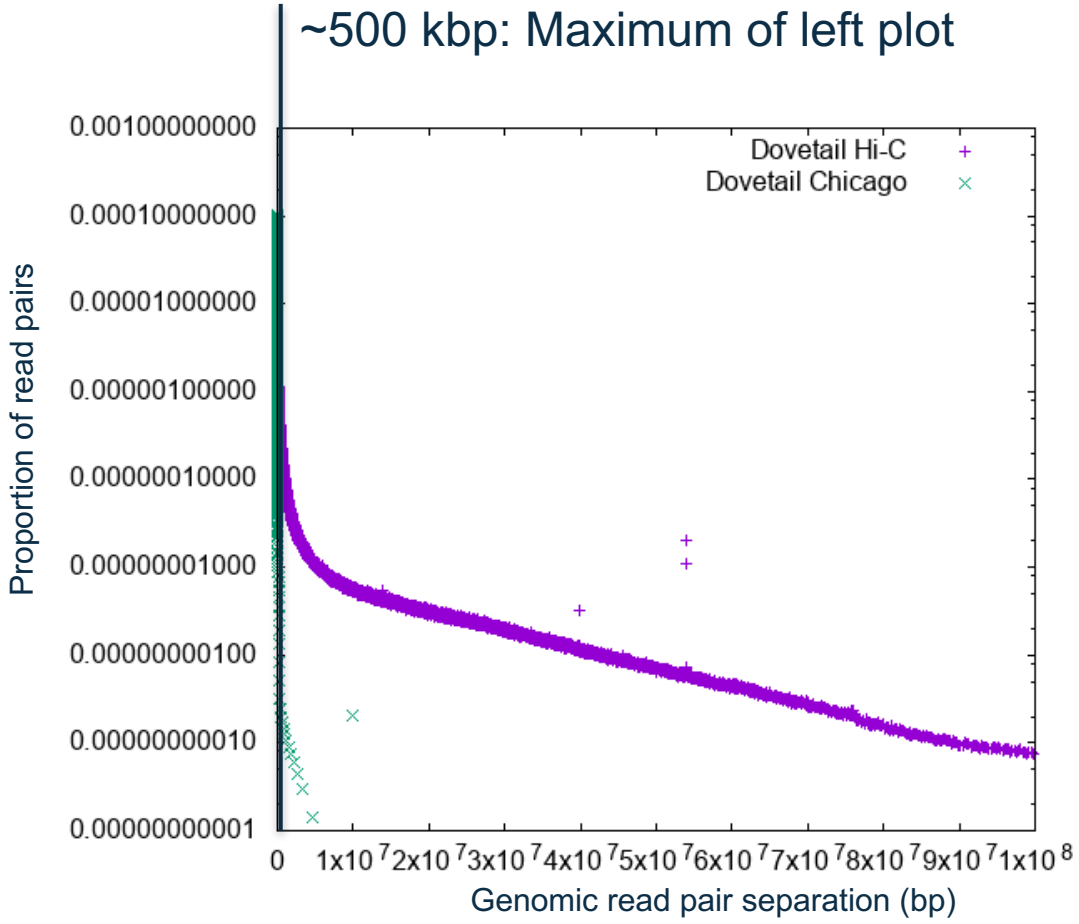



**Dovetail™ Hi-C** libraries start from tissue or cell-culture and endogenous chromatin is extracted after fixation.

**HiRise™** Scaffolding Pipeline

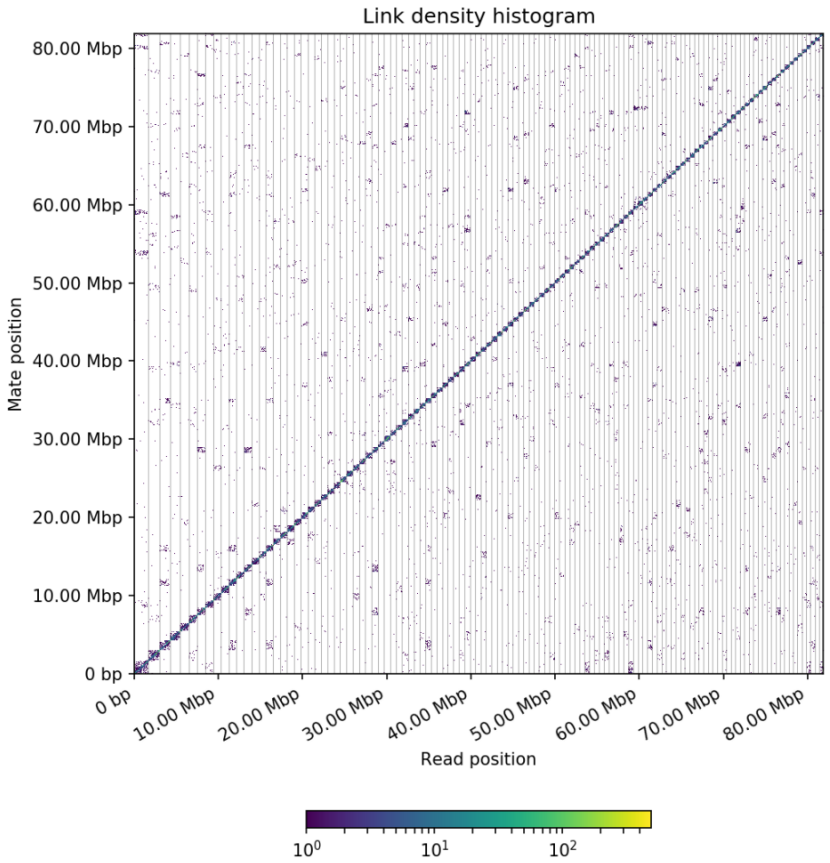
# Information Scales



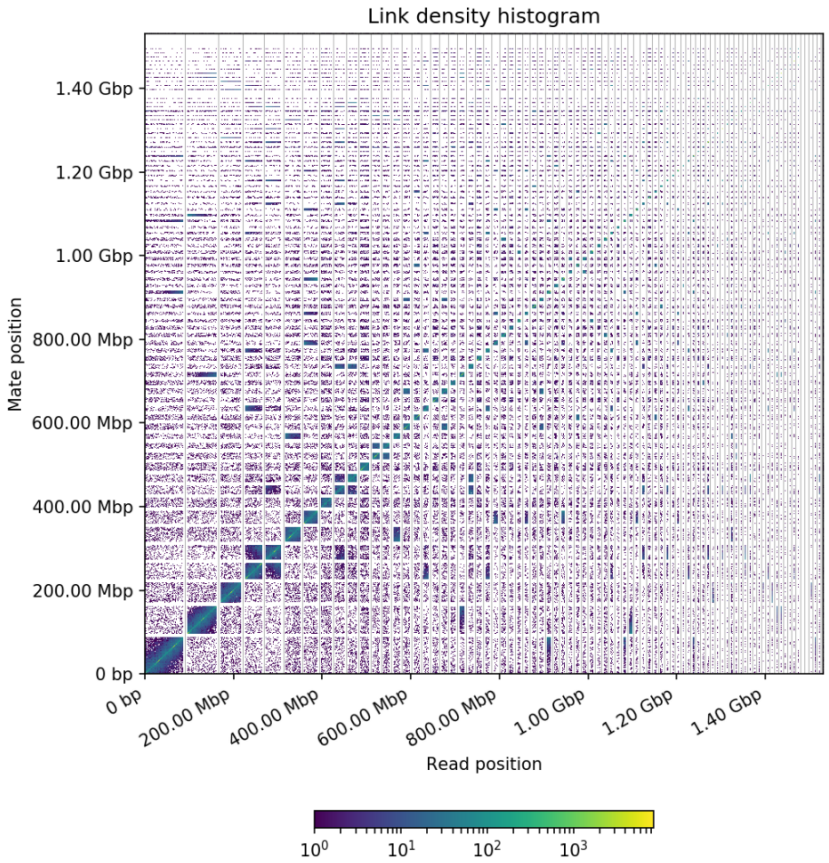



# Soup to Nuts

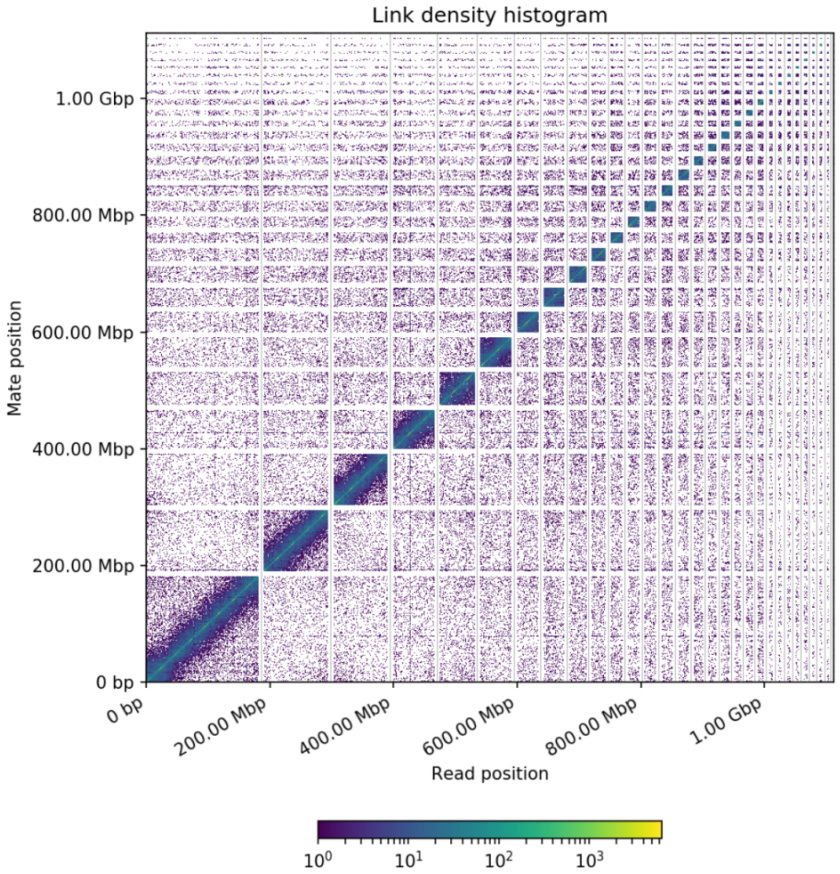
*de novo*




Chicago



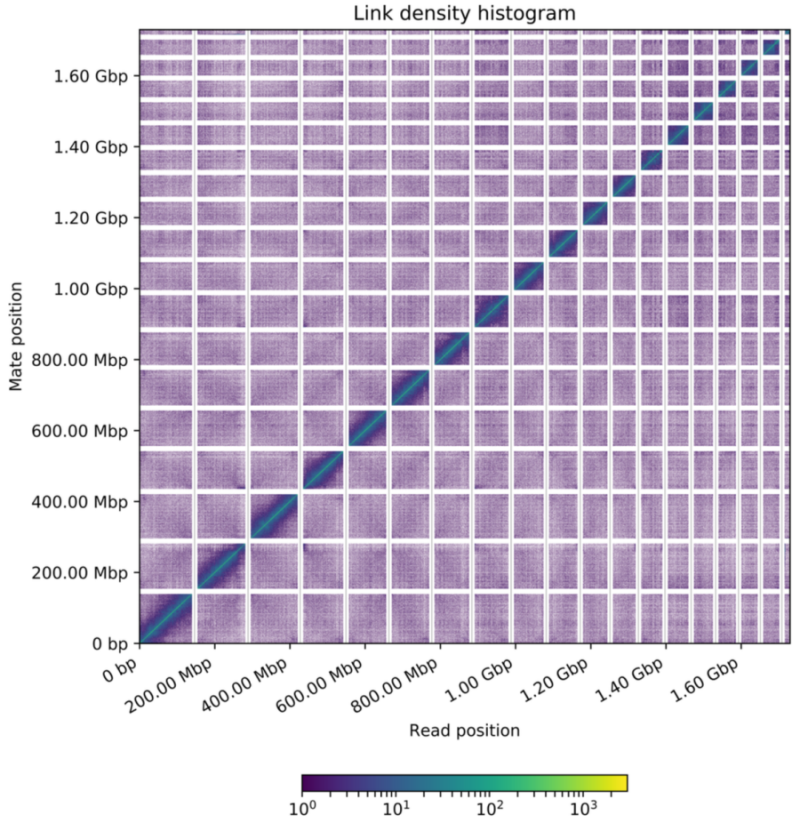

Dovetail Hi-C



# All Manner of Critters

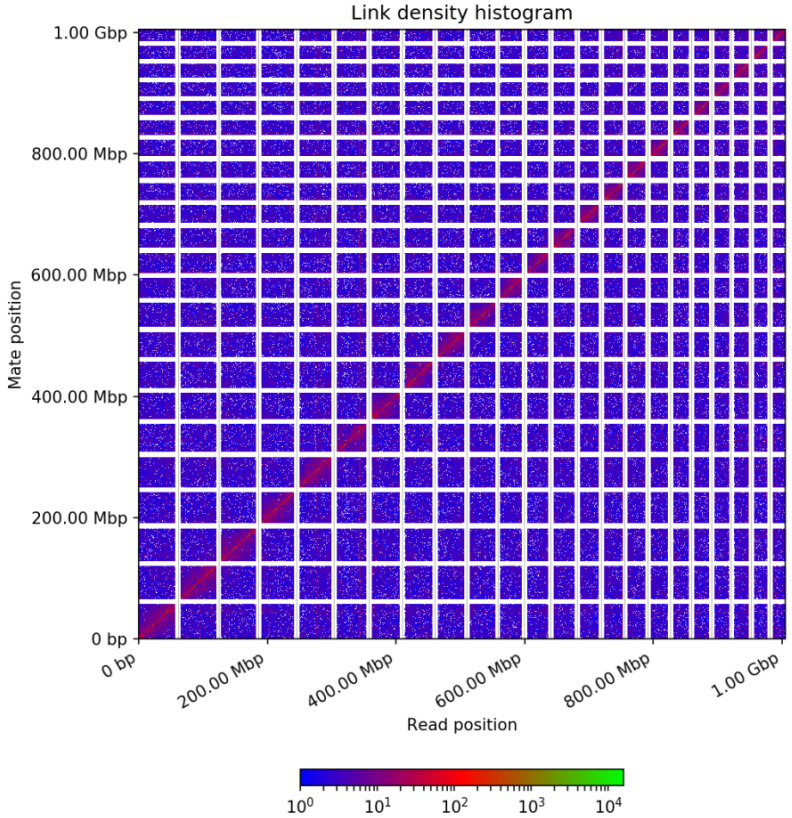
## Reptiles

N50: ~500 kbp > 15 Mbp > 90 Mbp



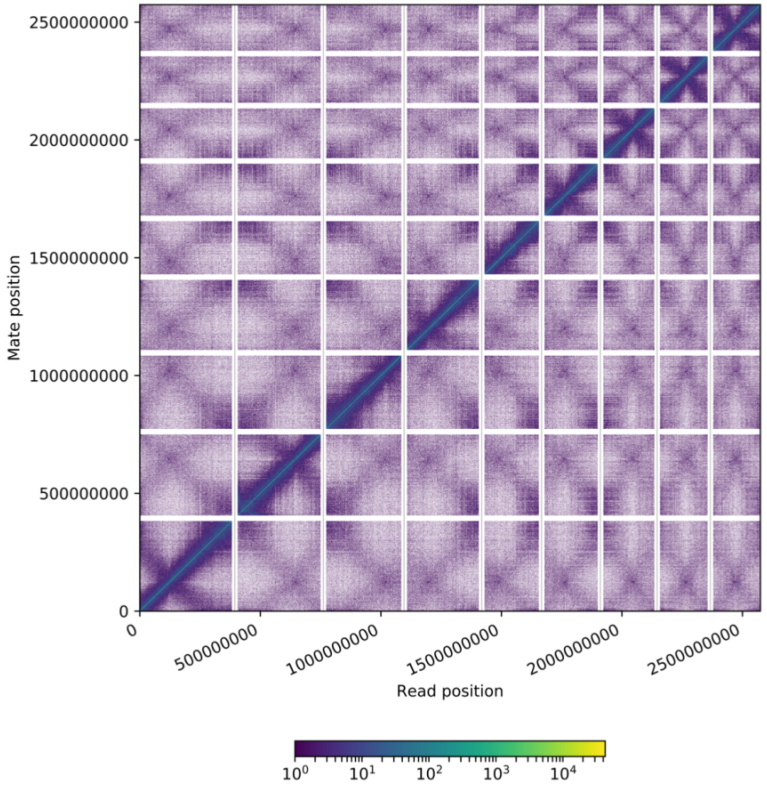

## Fish

24 kbp > 12 Mbp > 41 Mbp




## Plants

~500 kbp > 6.5 Mbp > 295 Mbp



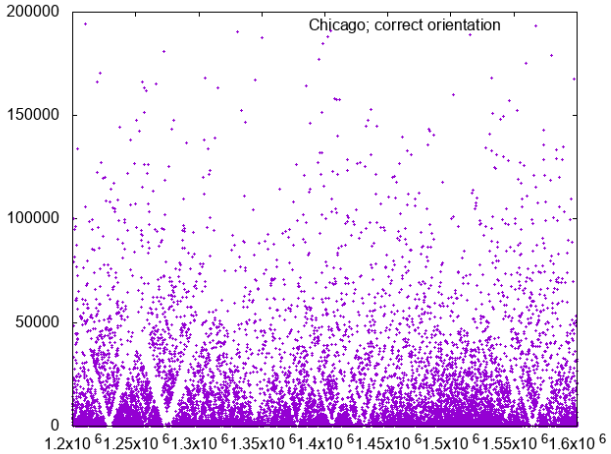
# Complementarity

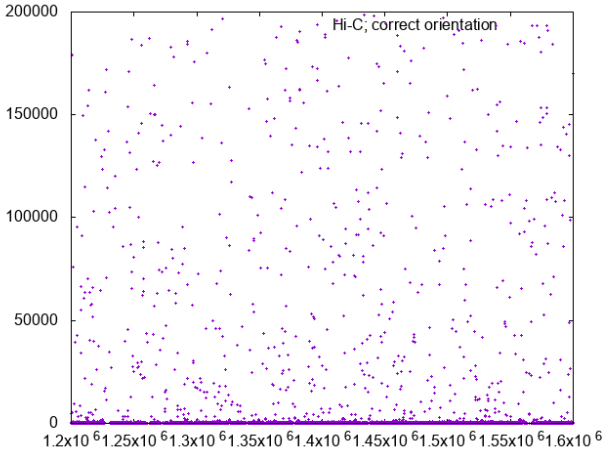
Chicago Data

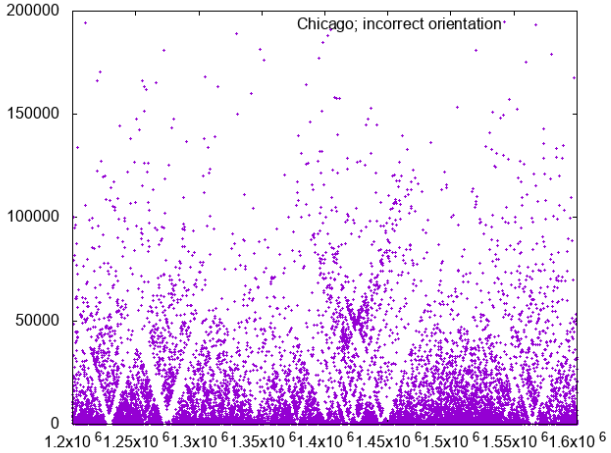
Hi-C Data

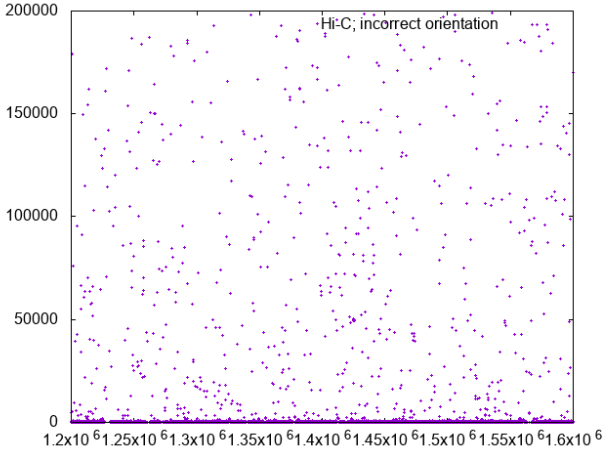
Correct Orientation
(Scaffold #1)

Incorrect Orientation
(Scaffold #2)




Dovetail
GENOMICS

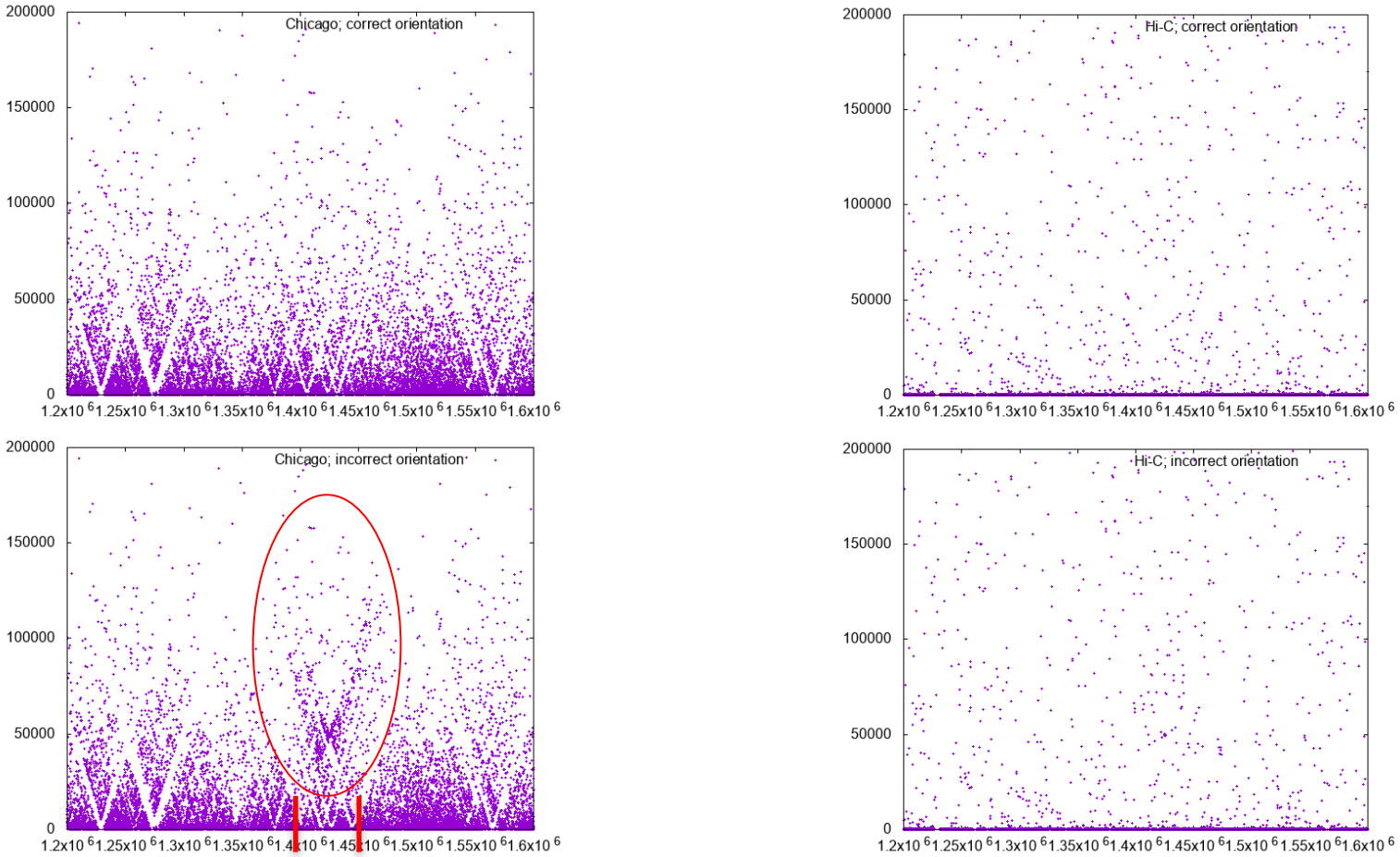
# Complementarity

Chicago Data

Hi-C Data

Correct Orientation
(Scaffold #1)

Incorrect Orientation
(Scaffold #2)




Dovetail
GENOMICS

# Reading through the repeats: Dovetail technology improves assembly of insect genomes

Brenda Oppert

USDA ARS Center for Grain and Animal Health Research

1515 College Ave.

Manhattan, KS 66502 USA

785-776-2780; brenda.oppert@ars.usda.gov

# Stored product pests – Why study?

- Stored product insect pests adversely affect grain in storage, milling facilities, warehouses, and even the consumer pantry
- The economic impact of grain production in the United States alone is estimated at $115 billion annually
- As grains are processed and manufactured into human and animal food products, their value increases dramatically at which point insect infestations further exacerbate loss
- It is estimated that insects destroy 5-10% of stored grain in developed countries and >30% in developing countries
- **The overall impact of storage pests is estimated to be as high as $300 trillion**

# Stored product pests – Why study?

- The ability to effectively manage storage pests has been challenged due to the loss of methyl bromide as a structural fumigant and **increasing insect resistance to the fumigant phosphine**
- Factors that elevate concerns include the roles insects may play as allergens and in the transport of pathogens, invasions of new pest species, and range expansion of other pests due to global climate change
- An increasing consumer demand for food that is organic or 'green' drives the development of non-chemical alternates
- The increasing oversight by both customers and federal inspection agencies has resulted in demands by the food industry for more effective integrative pest management (IPM) programs

# Functional genomics to understand pest biology and develop targeted management tools

- What can we learn from genetics about the success of insect pests in their environments?
- Can genetics inform us about how resistance to insecticides develops?
- Are there vulnerabilities in the insect genome that we can target with new control strategies (i.e., oral RNAi, enzyme inhibitors, etc)?
- How have different insect pests evolved over time to adapt to hostile environments?
- Can reproductive genes be targeted?
- Are gene drives feasible and desirable?

# We need sequenced genomes!

- First stored product insect with a sequenced genome was *Tribolium castaneum* (red flour beetle)
  - Also, the first beetle genome to be sequenced, and the first agriculturally-important insect
  - Accomplished entirely with Sanger Sequencing, inbred strain
  - *Tribolium* Genome Sequencing Consortium. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452: 949-955.
  - Next?

# *Rhyzopertha dominica* (lesser grain borer)

- The immature stage feeds within the kernel, making control more difficult
- Has rapidly developed resistance to a number of control products
- Obtained inbred line (more than 20 generations)
- Genome is estimated at 476 Mb
- Sequenced gDNA on PGM (12x), MiSeq (17x), HiSeq (1x), and PacBio (54x, P6 chemistry)
  - gDNA extracted with ZymoResearch kits
  - Tried adults and pupae – pupae worked best
  - Extracted from male and female separately pooled
  - Multiple (multiple) assemblies
    - SOAP deNovo (MiSeq)
    - mHAP and **CANU** (PacBio)
    - **Dovetail!**

**X-ray of *R. dominica* larvae feeding in grain kernels**

# Dovetail Assembly of *R. dominica*

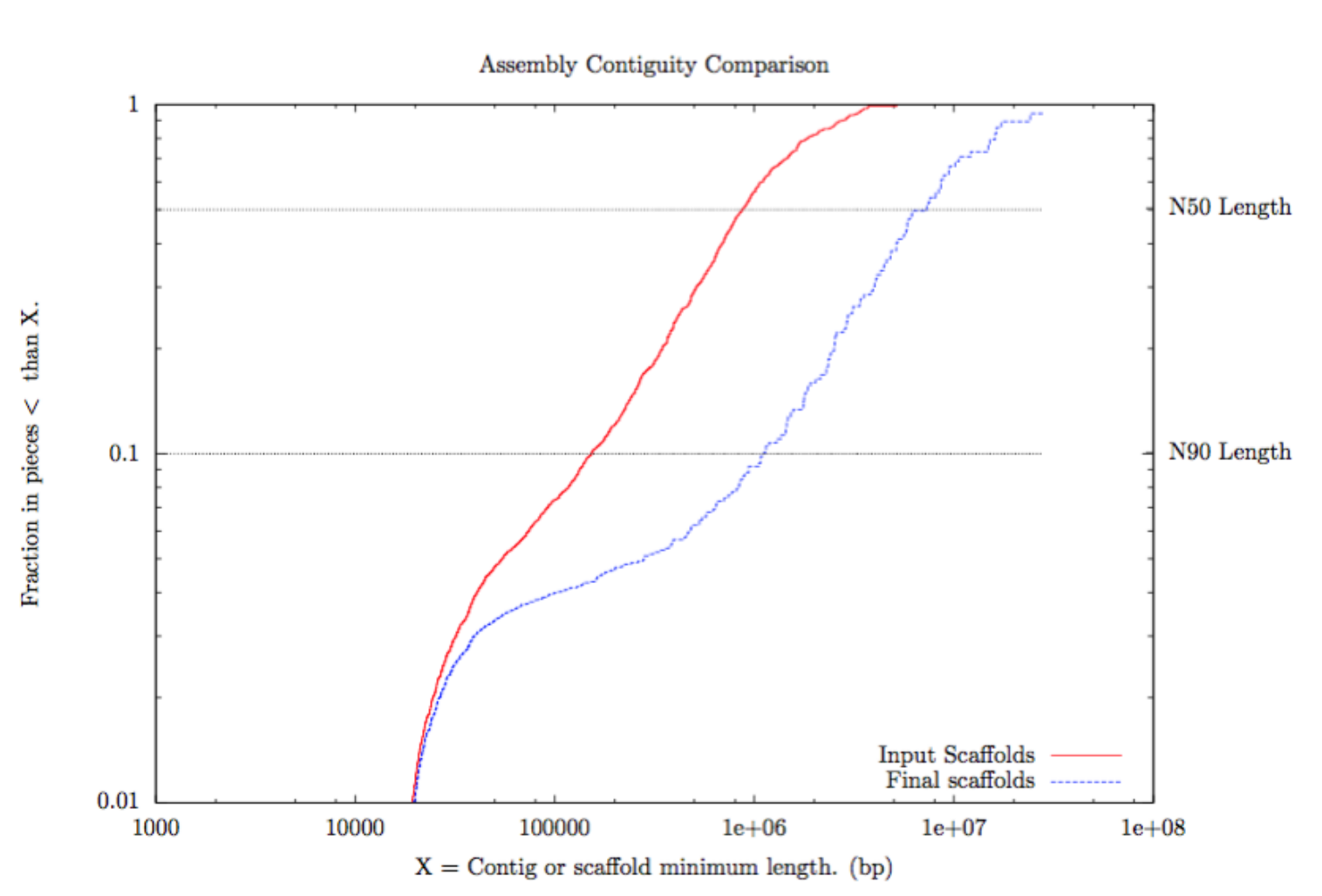
Estimated Chicago physical coverage (1-50 Kb pairs): 133.6 X

| | Starting Assembly | Dovetail HiRise Assembly |
|---|---|---|
| Total Length | 493.3 Mb | 493.4 Mb |
| N50 Length | 158 scaffolds ; min length 0.871 Mb | 20 scaffolds ; min length 7.32 Mb |
| N90 Length | 627 scaffolds ; min length 0.153 Mb | 84 scaffolds ; min length 1.11 Mb |

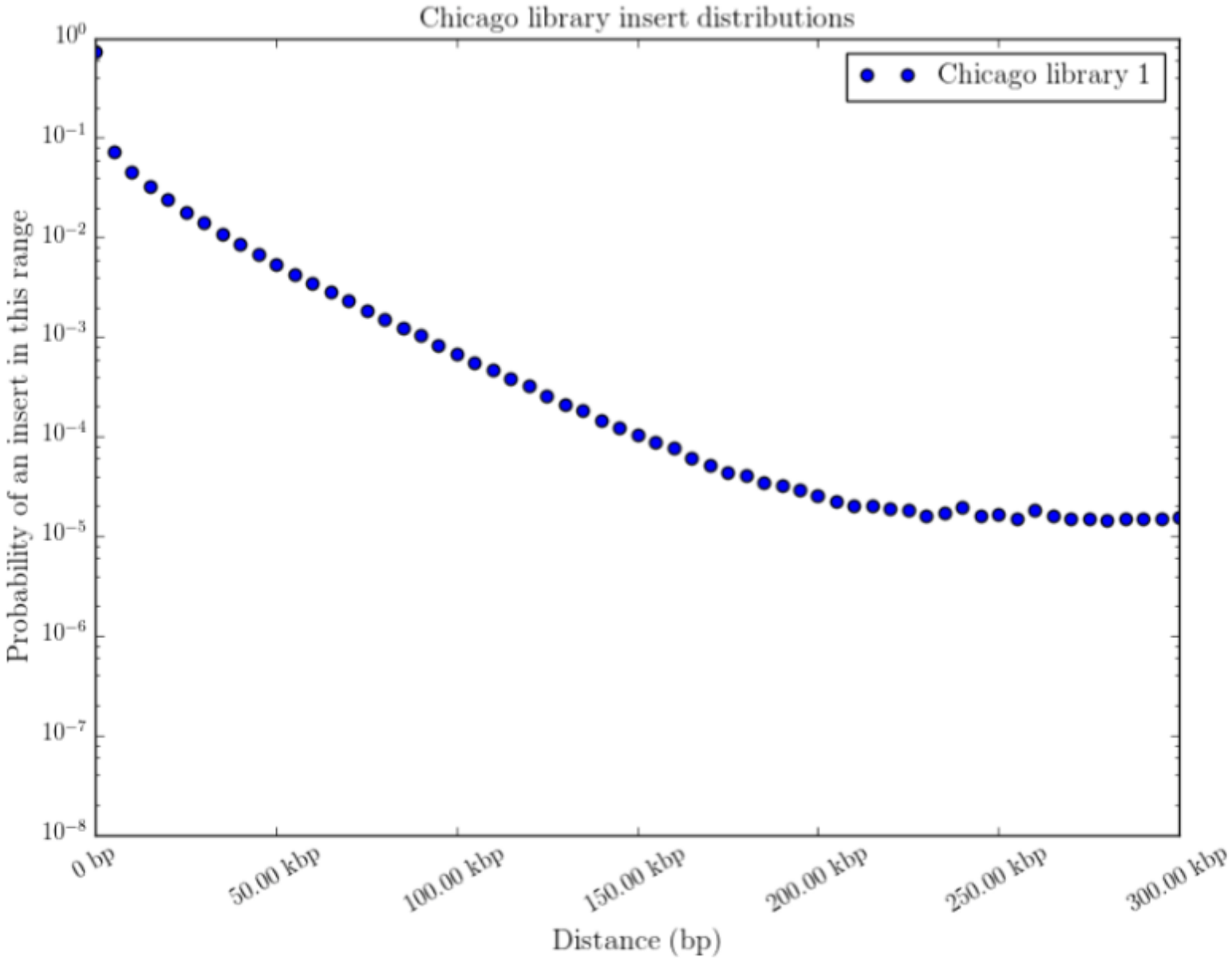
| Comparative Assembly Statistics | | |
|---|---|---|
| | Input Assembly | Dovetail HiRise Assembly |
| Longest scaffold | 5205710 | 27933969 |
| Number of scaffolds | 1861 | 948 |
| Number of scaffolds >1 kb | 1861 | 948 |
| Contig N50 | 871.5 kb | 860.8 kb |
| Number of gaps | 0 | 968 |
| Percent of genome in gaps | 0% | 0.02% |

* Note: Every join made by HiRise creates a gap.

| Other Statistics | |
|---|---|
| Number of breaks made to input assembly by HiRise | 55 |
| Number of joins made by HiRise | 968 |
| Chicago library 1 stats | 53M read pairs; 2x101 bp |




A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. Y-axis: the fraction of the assembly; X-axis: scaffold length (bp). The two dashed lines mark the N50 and N90 lengths of each assembly. This plot excludes scaffolds less than 1 kb.




This figure shows the distribution of insert sizes in the Chicago library. The distance between the forward and reverse reads is given on the X-axis in basepairs, and the probability of observing a read pair with a given insert size is shown on the Y-axis.

# Post Dovetail

- SeqManNGen (DNAStar) for hybrid assembly of MiSeq and DT contigs

| Assembly | # Contigs | Total bp | N50 | Median Length | Average Length | Standard Length |
|---|---|---|---|---|---|---|
| CANU | 1,861 | 493,284,530 | 871,450 | 47,136 | 265,064 | 490,388 |
| Dovetail | 948 | 493,381,330 | 7,324,187 | 23,002 | 520,444 | 2,110,438 |
| Hybrid | 336 | 479,149,650 | 7,435,960 | 64,277 | 1,426,041 | 3,361,090 |

- BUSCO* metrics (Insecta)

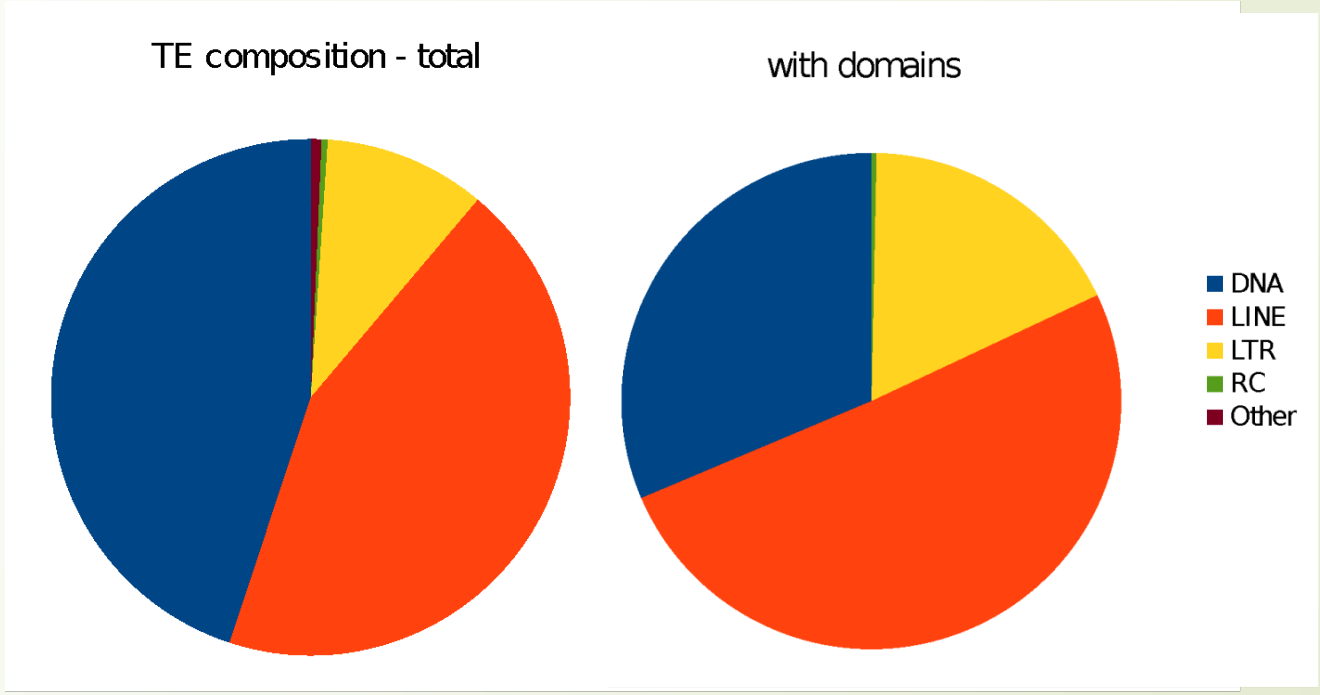
| Assembly | Complete | Single | Duplicate | Fragment | Missing |
|---|---|---|---|---|---|
| CANU | 99.3 | 98.5 | 0.8 | 0.5 | 0.2 |
| Dovetail | 99.4 | 98.6 | 0.8 | 0.3 | 0.3 |
| Hybrid | 97.6 | 97.0 | 0.6 | 0.5 | 1.9 |

- Augustus – 37,208 genes, masked – 22,925

**BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.*Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov*Bioinformatics*, published online June 9, 2015

# *R. dominica* repeats – Highly diverse TE landscape

- Highly repetitive genome, common for arthropods
- ~36% transposon-related sequences
- **Only ~9% are similar to typical transposons**
- Most are simple repeats or remnants of ancient transposons
- Size of genome suggests genome expansion by mobile element proliferation
  - Low GC content (35%)
  - Repeats usually AT rich found in AT rich regions



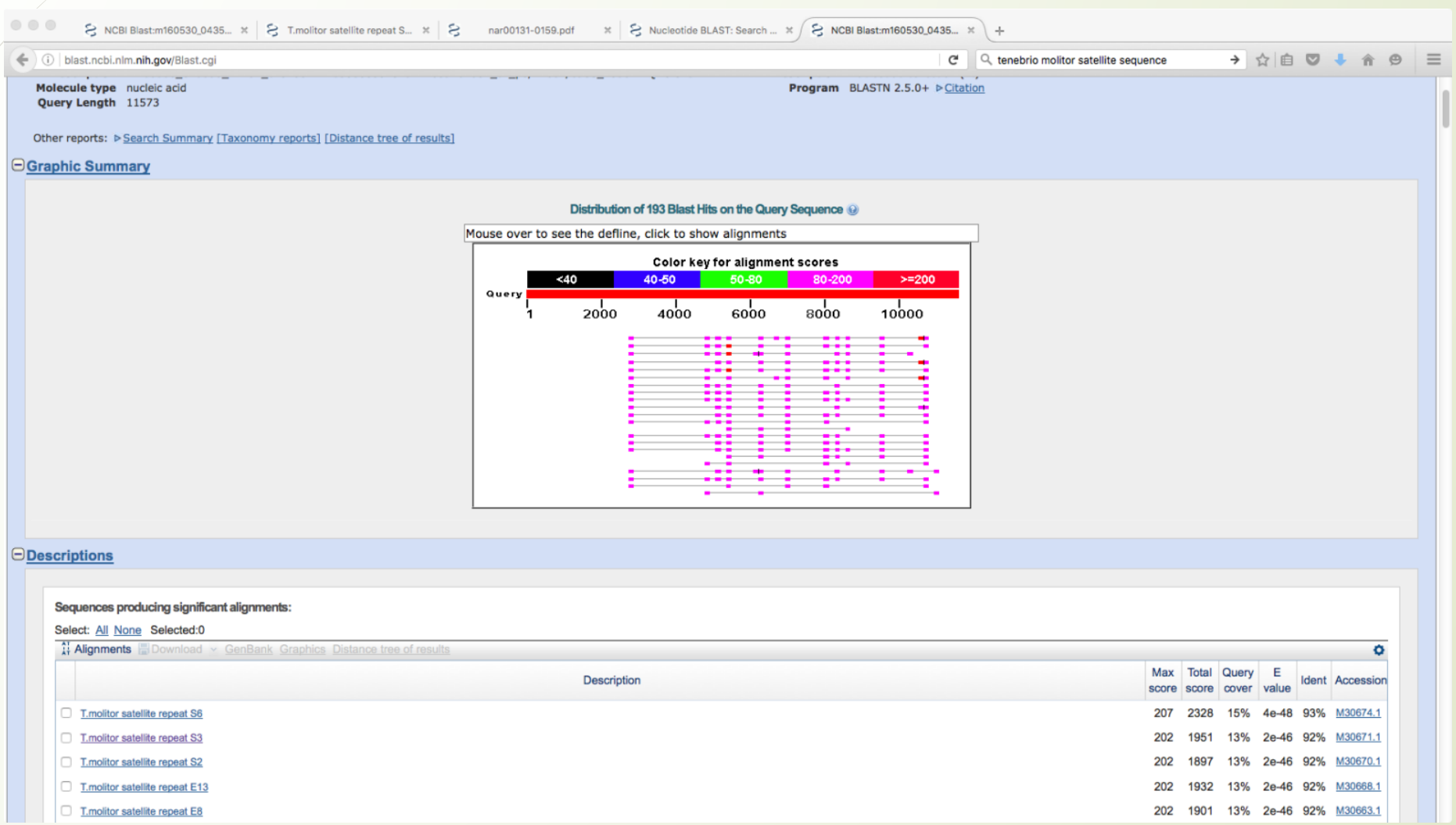

- Mostly LINE (Penelope, L2, and CR1) and LTR retrosposon (Ty3/Gypsy) families
- **More than 119k remnant copies and 7k copies Tc1/Mariner superfamily**
- DNA transposons with DDE transposases from 15 superfamilies
- A variety of less common TE
  - LTR DIRS elements
  - Cryptons and rolling circle Helitrons
  - A wide variety of remnants of many DNA and LINE families

Anna Muszewska, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Pawinskiego 5A, 02-106 and Kamil Steczkiewicz, Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Zwirki i Wigury 93, 02-089

# *Tenebrio molitor* (yellow mealworm)

- More problematic in places like poultry houses, but also is a commodity as larvae are sold as reptile and bird feed, also food source in some parts of the world
- Larger in size and has a longer developmental time, taking up to two years to complete development in the field (4-6 months in the lab)
- Model for biochemical studies of the coleopteran gut for more than 20 years, and we know a great deal about how larvae digest food
- Genome is 509 Mb
- Project began as proof of concept, but is now the basis for a collaboration with All Things Bugs (Dr. Aaron Dossey) and NCSU (Dr. Marce Lorenzen)
- gDNA extracted from a single male pupa
  - Best extraction for long gDNA was with Omega E.Z.N.A. Tissue DNA kit http://omegabiotek.com/store/product/e-z-n-a-tissue-dna-kit
- Sequenced gDNA on MiSeq (56x) and PacBio (45x, P6 chemistry)
  - Assembly with CANU (PacBio)
    - 1st attempt overwhelmed the data storage, writing terabytes of data and days and days of compute time
    - 2nd attempt by CANU developer Sergey Koren (NIH) also had problems, but after discovering the problem, customized CANU to suppress repeats during assembly. Running under normal parameters estimated to take 20k CPU hours, similar to human genome
  - **Dovetail!**

# Repeat satellites in *T. molitor*




| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| T.molitor satellite repeat S6 | 207 | 2328 | 15% | 4e-48 | 93% | M30674.1 |
| T.molitor satellite repeat S3 | 202 | 1951 | 13% | 2e-46 | 92% | M30671.1 |
| T.molitor satellite repeat S2 | 202 | 1897 | 13% | 2e-46 | 92% | M30670.1 |
| T.molitor satellite repeat E13 | 202 | 1932 | 13% | 2e-46 | 92% | M30668.1 |
| T.molitor satellite repeat E8 | 202 | 1901 | 13% | 2e-46 | 92% | M30663.1 |

# Repeat satellites in *T. molitor*

- Davis, CA and Wyatt, GR. 1989. Distribution and sequence homogeneity of an abundant satellite DNA in the beetle, *Tenebrio molitor*. Nuc. Acids Res. 17:5579.
  - Comprises up to 60% of genome
  - Present in all chromosomes
  - 142 nt, less than 2% divergence in sequence among 18 different satellites
  - Suggests that homogeneity of the satellite is preserved
  - A nightmare for assembly algorithms!

# Dovetail Assembly of *T. molitor*

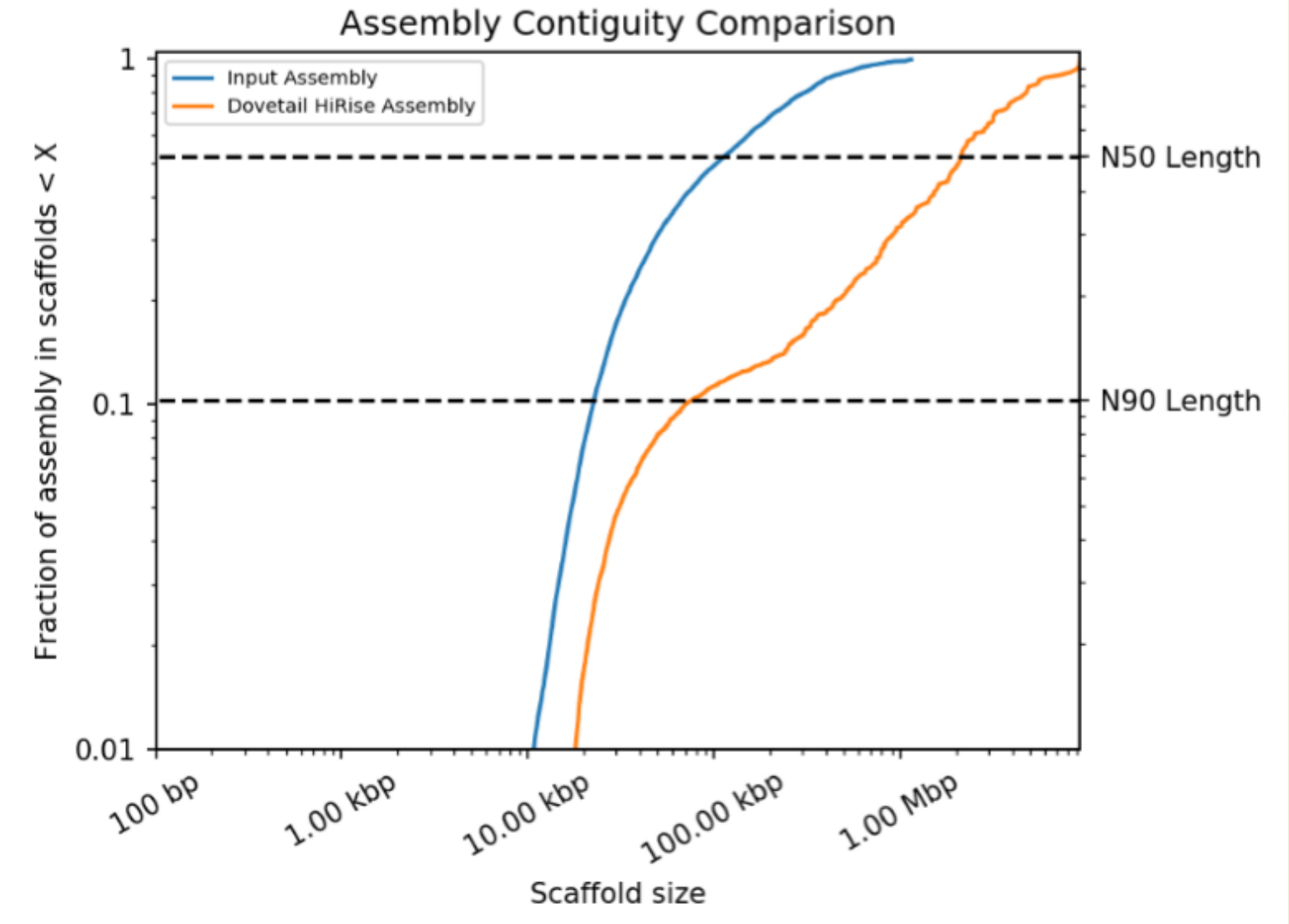
Estimated physical coverage (1-100 kb pairs): 53.30X

| | Input Assembly | Dovetail HiRise Assembly |
|---|---|---|
| Total Length | 417.68 Mb | 423.05 Mb |
| L50/N50 | 908 scaffolds; 0.104 Mb | 58 scaffolds; 2.013 Mb |
| L90/N90 | 4,686 scaffolds; 0.023 Mb | 476 scaffolds; 0.057 Mb |

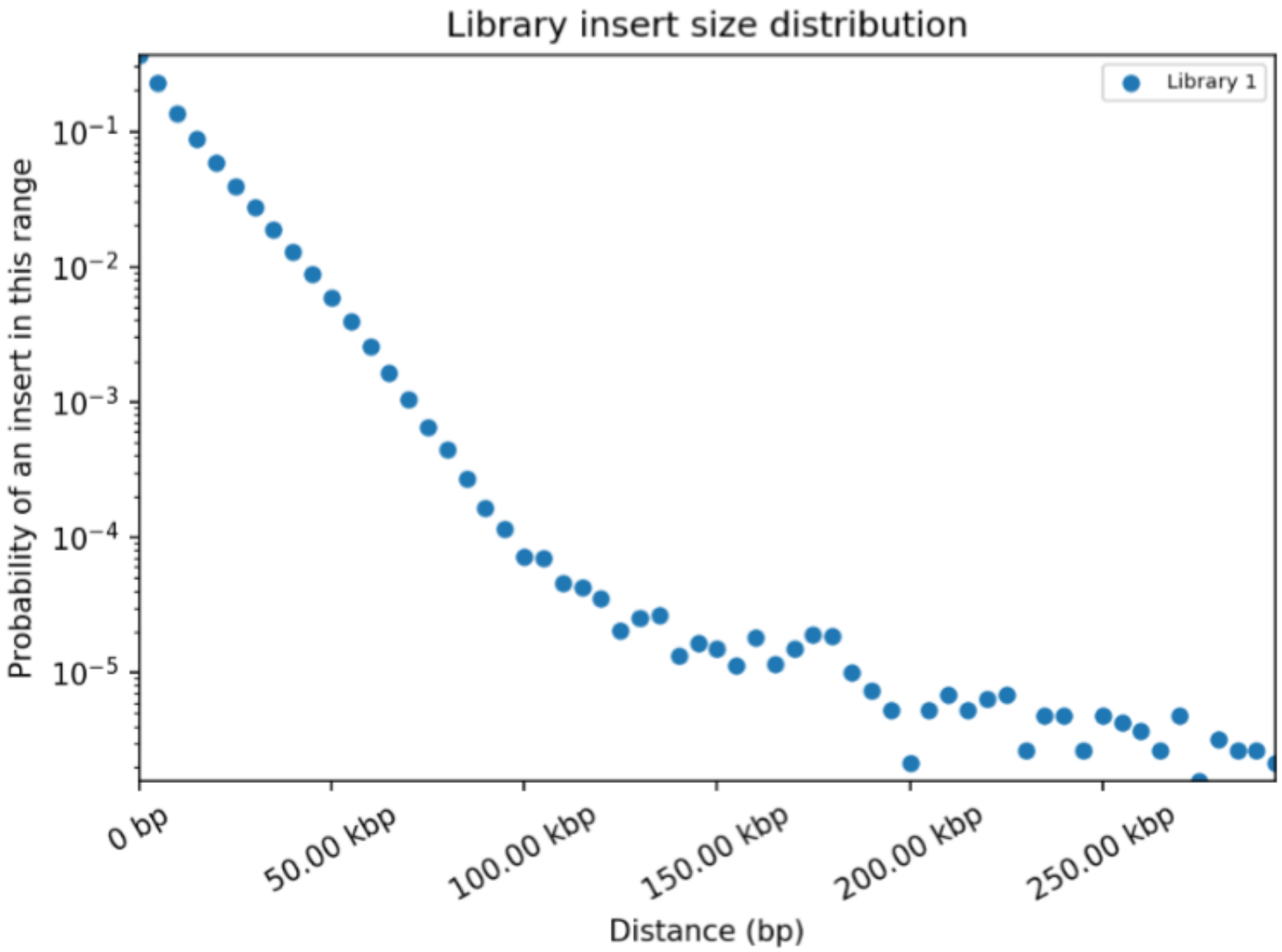
| Comparative Assembly Statistics | | |
|---|---|---|
| | Input Assembly | Dovetail HiRise Assembly |
| Longest Scaffold | 1,144,088 bp | 9,264,513 bp |
| Number of scaffolds | 7,484 | 2,364 |
| Number of scaffolds > 1kb | 7,484 | 2,364 |
| Contig N50 | 103.70 kb | 92.55 kb |
| Number of gaps | 0 | 5,376 |
| Percent of genome in gaps | 0.00% | 1.27% |

* Note: Every join made by HiRise creates a gap.

| Other Statistics | |
|---|---|
| Number of breaks made to input assembly by HiRise | 256 |
| Number of joins made by HiRise | 5,376 |
| Number of gaps closed after HiRise | 0 |
| Library 1 stats | 169M read pairs; 2x151 bp |




A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. Y-axis: the fraction of the assembly; X-axis: scaffold length (bp). The two dashed lines mark the N50 and N90 lengths of each assembly. This plot excludes scaffolds less than 1 kb.




This figure shows the distribution of insert sizes in the Chicago library. The distance between the forward and reverse reads is given on the X-axis in basepairs, and the probability of observing a read pair with a given insert size is shown on the Y-axis.

# New to Dovetail reports

| BUSCO Stats | | | | | |
|---|---|---|---|---|---|
| | **Single copy** | **Duplicated** | **Fragmented** | **Missing** | **Total** |
| **Input Assembly** | 253 | 37 | 4 | 9 | 303 |
| **Dovetail HiRise Assembly** | 259 | 28 | 6 | 10 | 303 |

Number of BUSCO (Benchmarking Universal Single-Copy Ortholog) genes found in the assembly before and after HiRise using the eukaryota odb9 dataset. Genes are split into four categories: complete and single-copy, complete and duplicated, fragmented, and missing.

# Post Dovetail

- SeqManNGen (DNAStar) for hybrid assembly of MiSeq and DT contigs

| Assembly | # Contigs | Total bp | N50 | Median Length | Average Length | Standard Length |
|---|---|---|---|---|---|---|
| CANU | 7,484 | 417,676,750 | 103,701 | 28,448 | 55,809 | 85,355 |
| Dovetail | 2,364 | 423,052,750 | 2,013,304 | 23,820 | 178,956 | 667,558 |
| Hybrid | 1,293 | 400,737,566 | 2,120,994 | 36,138 | 309,929 | 887,069 |

- BUSCO* metrics (Insecta)

| Assembly | Complete | Single | Duplicate | Fragment | Missing |
|---|---|---|---|---|---|
| CANU | 96.1 | 82.7 | 13.4 | 1.2 | 2.7 |
| Dovetail | 96.2 | 86.7 | 9.5 | 1.2 | 2.6 |
| Hybrid | 96.3 | 84.0 | 12.3 | 0.8 | 2.9 |

- Augustus – 49,171 genes, masked in progress

**BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.*Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov*Bioinformatics*, published online June 9, 2015

# Next Up

***S. oryzae* pupa in a peeled-back kernel**

- *Sitophilus oryzae* (rice weevil)
  - Also an internal feeder
  - Doesn't just attack rice, but also other cereals such as wheat and maize
  - Weevils are major pests in agriculture, and the cotton industry has battled a related beetle, the boll weevil
  - No genomic or transcriptome data for *Sitophilus* spp., but inbreeding was initiated last year and we have single pair mated lines for +ten generations for a genome sequencing project
  - PacBio sequencing in progress (collaborator Tim Smith, USDA ARS US Meat Animal Research Center, Clay Center, NE)
  - Genome size estimated at 770 Mb

# Summary

- Dovetail has greatly reduced the number of contigs (2- to 3.2-fold reduction)and increased the N50 (8.4- to 19.4 increase!) in our insect genomes
- SeqManNGen with MiSeq data (hybrid assembly) further improved metrics
- BUSCO scores indicate that the two insect genomes are nearly complete in sequencing of genes
- Dovetail technology works particularly well with complex, repetitive genomes
- Next test will be *Sitophilus oryzae*, with approximately 2x genome size, and no information on repetitive sequences

# Questions regarding repetitive sequences

- Why are they retained and expanded in insect genomes?
- What is the evolutionary significance?
- Do they participate in the 3-D stability of chromosomes?
- Are they involved in gene regulation?
- Part of vehicles carrying tandem repeats?
  - Evidence of TE near gene expansion groups

# Collaborators

- ARS
  - Jeff Lord and **Kris Hartzer**
  - **Ken Friesen and Tom Morgan**
  - **Tim Smith** and Christy Kelley
  - Erin Scully and Rob Morrison
  - James Campbell
- NIH
  - **Sergey Koren** and Adam Phillippy
- All Things Bugs
  - Aaron Dossey
- University of Vermont
  - Yolanda Chen (Fanslo
- NCSU
  - Marce Lorenzen
- PacBio
  - **Richard Hall**
- Nimbix
  - Steve Hebert, Stephen Fox
- Australia
  - David Schlipalius, U Queensland
- Croatia
  - Miroslav Pohl, Ruder Boskovic´ Institute
- Poland
  - **Anna Muszewska**, Polish Academy of Sciences
  - Kamil Steczkiewicz, U Warsaw
- Russia
  - Elena Elpidina, Moscow State U
  - Alexander Martynov, Skoltech Center for Data-Intensive Biomedicine and Biotechnology (currently at MIT)
- DOVETAIL!
  - **Ed Green, Brandon Rice, Margot Hartley, Mark Daly**

Thank you!

Questions?