inary Draft Pange
Soldier Fly *Her*
Structural and
rom Diverse Po

r Rosche-Flores & Christine J P

diana University Indianapol

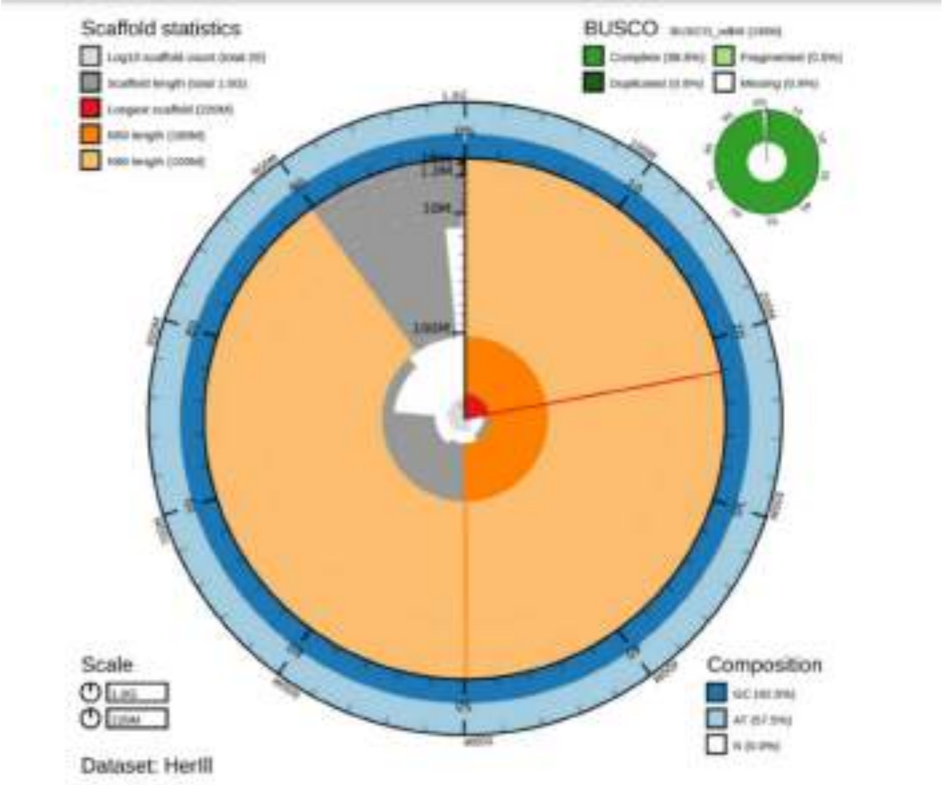
Black soldier fly draft pangenome

Broad Goal

- Download publicly available genomes
- Collect additional wild and domesticated samples for assembly
- Use the above to generate a pangenomic resource to support future research

Why do we need a pangenome?

1. The current *Hermetia illucens* reference established from a domesticated line



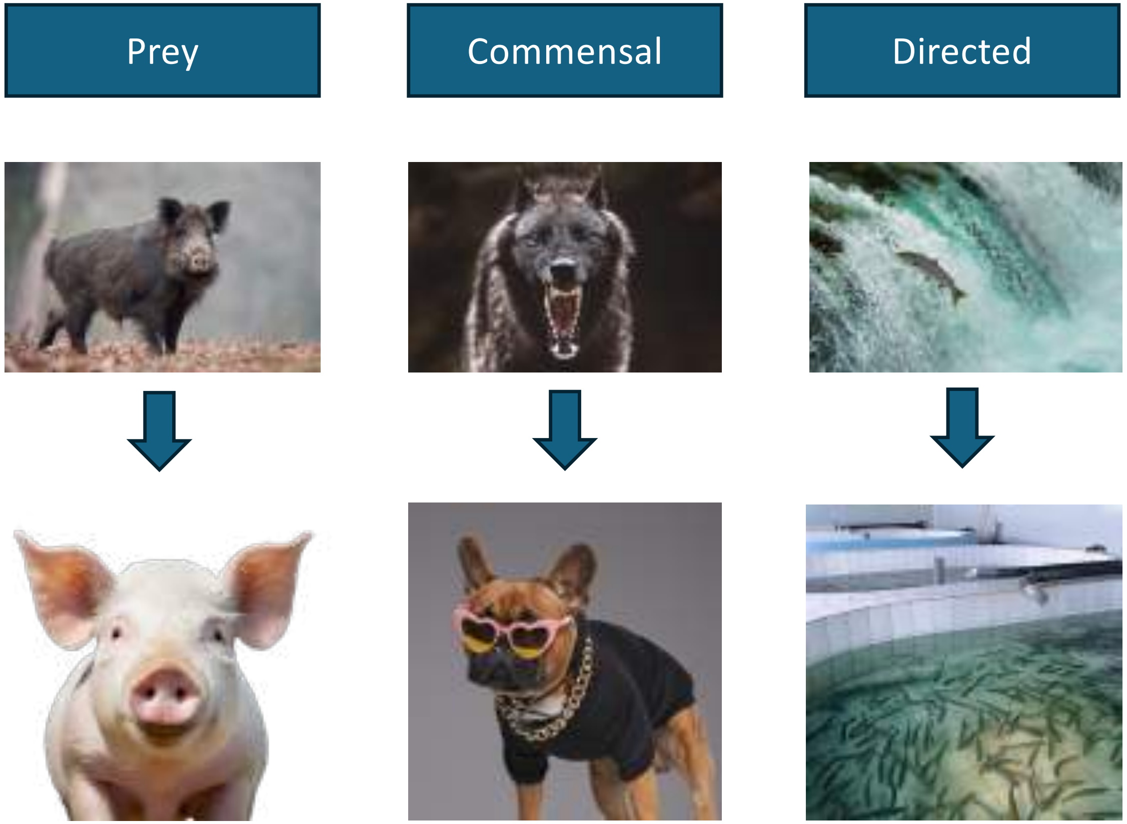
Generalovic *et al.*, 2021, *G3*

Domestication Changes Organisms

Prey: humans hunted animals, eventually captured and reared.

Commensal: animals attracted to humans for food, safety, etc.

Directed: humans target species directly for domestication



Prey

Commensal

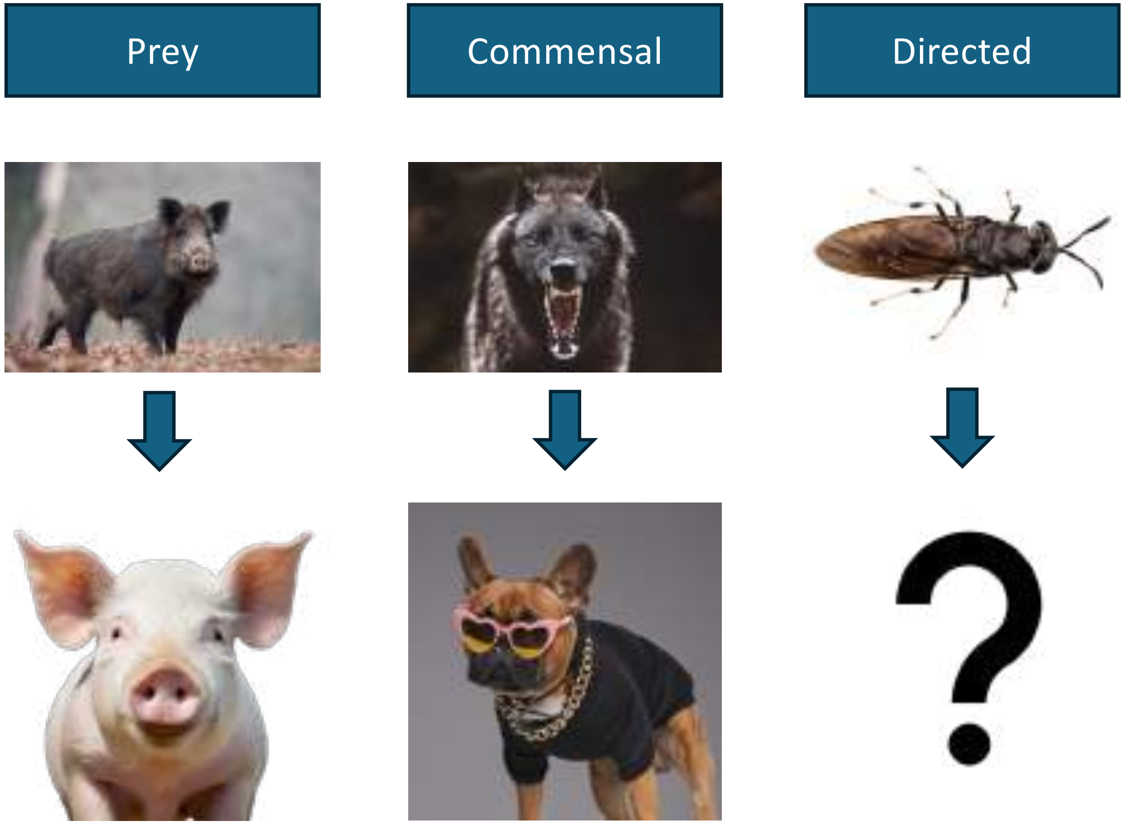
Directed

Domestication Changes Organisms

Prey: humans hunted animals, eventually captured and reared.

Commensal: animals attracted to humans for food, safety, etc.

Directed: humans target species directly for domestication



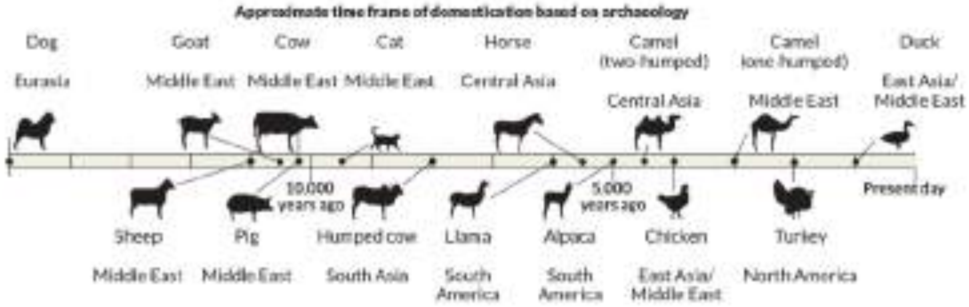
Prey

Commensal

Directed

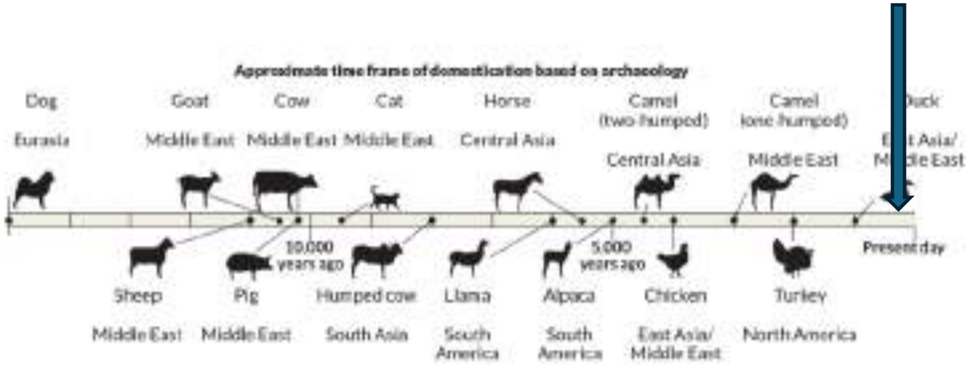
Directed Path

- Modern breeding programs



Directed Path

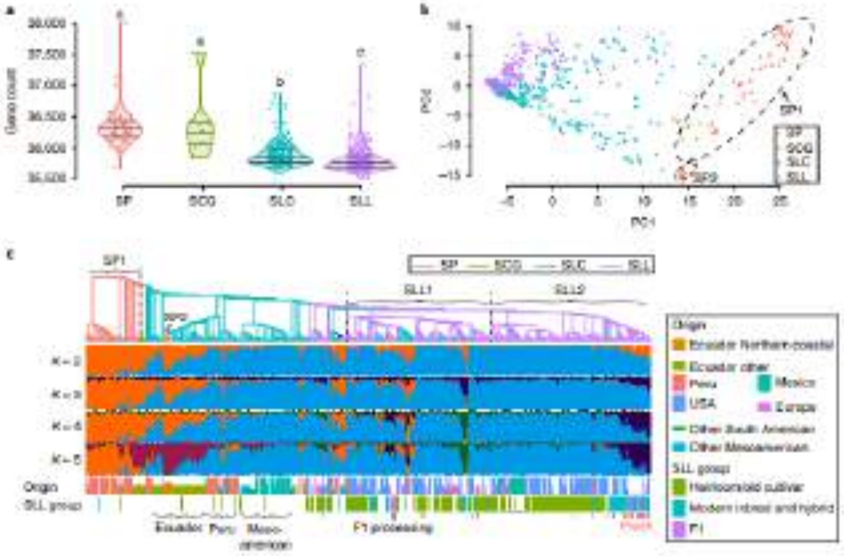
- Modern breeding programs
- Strong selection in short time: years vs thousands of years (dogs, cattle, silkworm)



https://www.collectedny.org/frameworkposts/timeline-of-animal-domestication/

Directed Path

- Modern breeding programs
- Strong selection in short time: years vs thousands of years (dogs, cattle, silkworm)
- Commercial relevant traits present in the wild may have been lost during this strong selection
- Ex: tomato flavor (Gao *et al*., 2019, *nature genetics*)



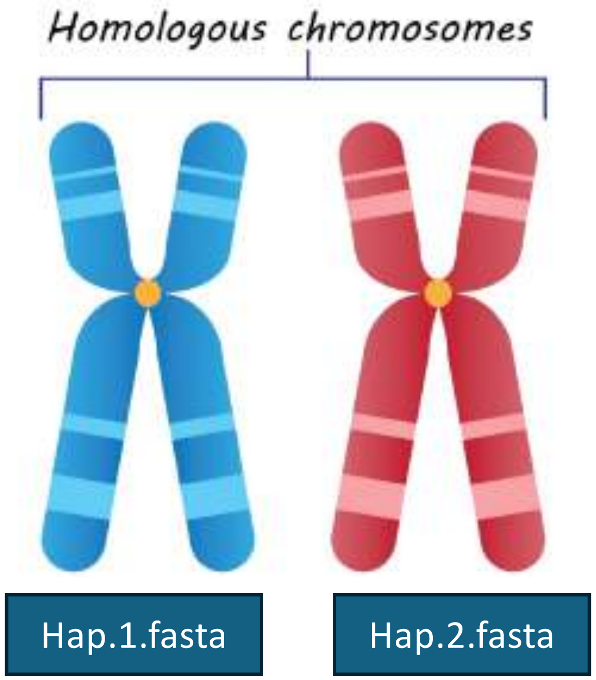
Gao *et al*., 2019, *nature genetics*

Why do we need a pangenome?

1. The current *Hermetia illucens* reference established from a domesticated
2. A single individual is not representative of a global population

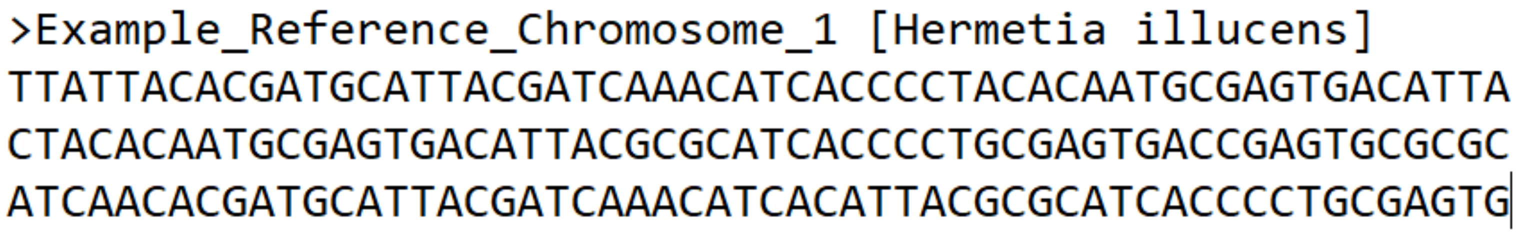
Reference Genomes

- All linear reference genomes suffer this limitation
- Fasta format does not natively represent diploids



Hap.1.fasta

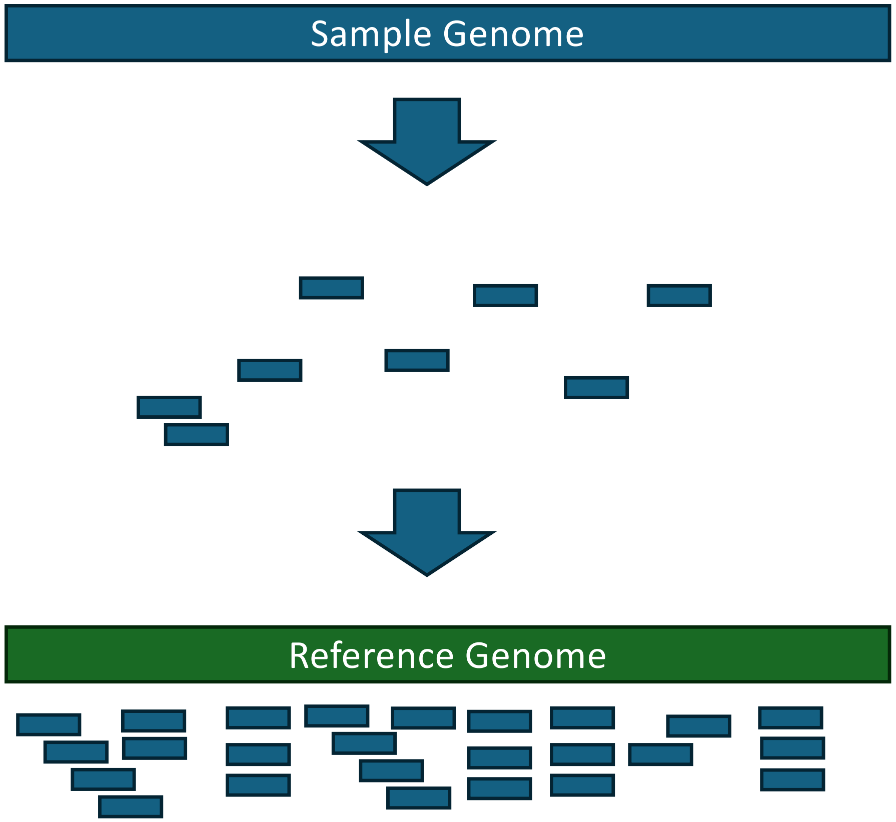
Hap.2.fasta



```
>Example_Reference_Chromosome_1 [Hermetia illucens]
TTATTACACGATGCATTACGATCAAACATCACCCCTACACAATGCGAGTGACATTA
CTACACAATGCGAGTGACATTACGCGCATCACCCCTGCGAGTGACCGAGTGCGCGC
ATCAACACGATGCATTACGATCAAACATCACATTACGCGCATCACCCCTGCGAGTG
```

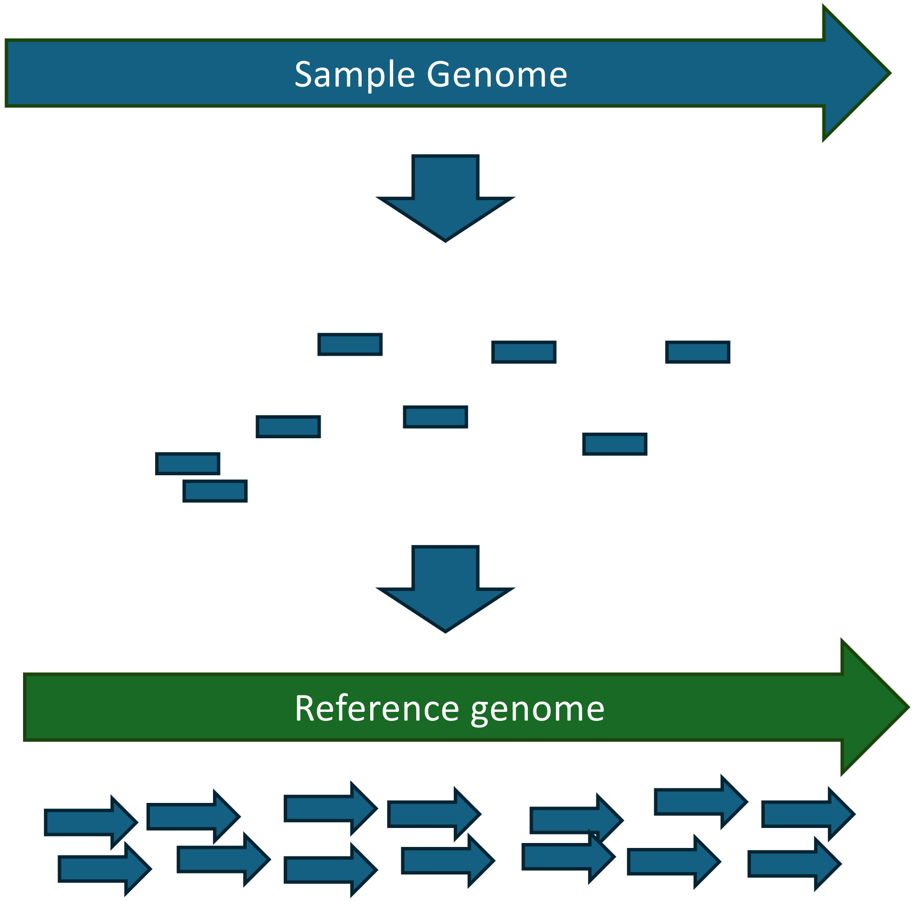
Reference Genomes

- Traditional methods involve mapping samples reads to reference



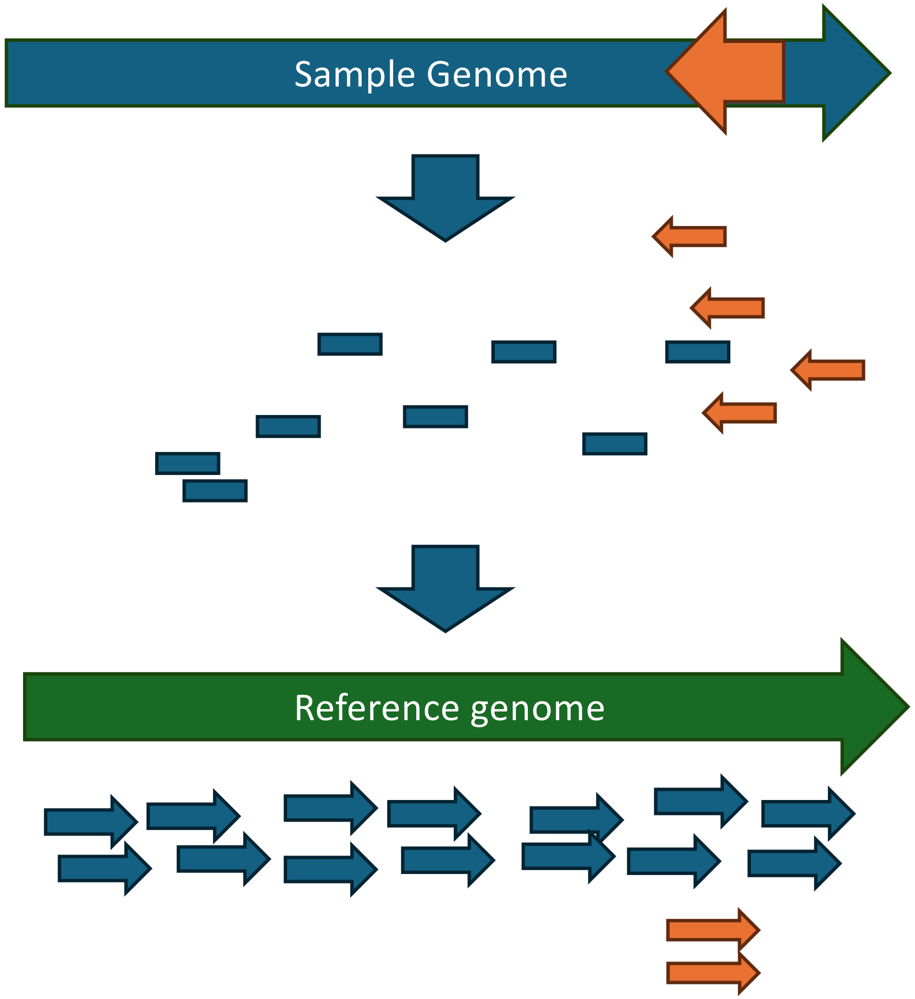
Reference Genomes

- Traditional methods involve mapping sample reads to reference
- Reference is assumed to be ground truth for orienting reads



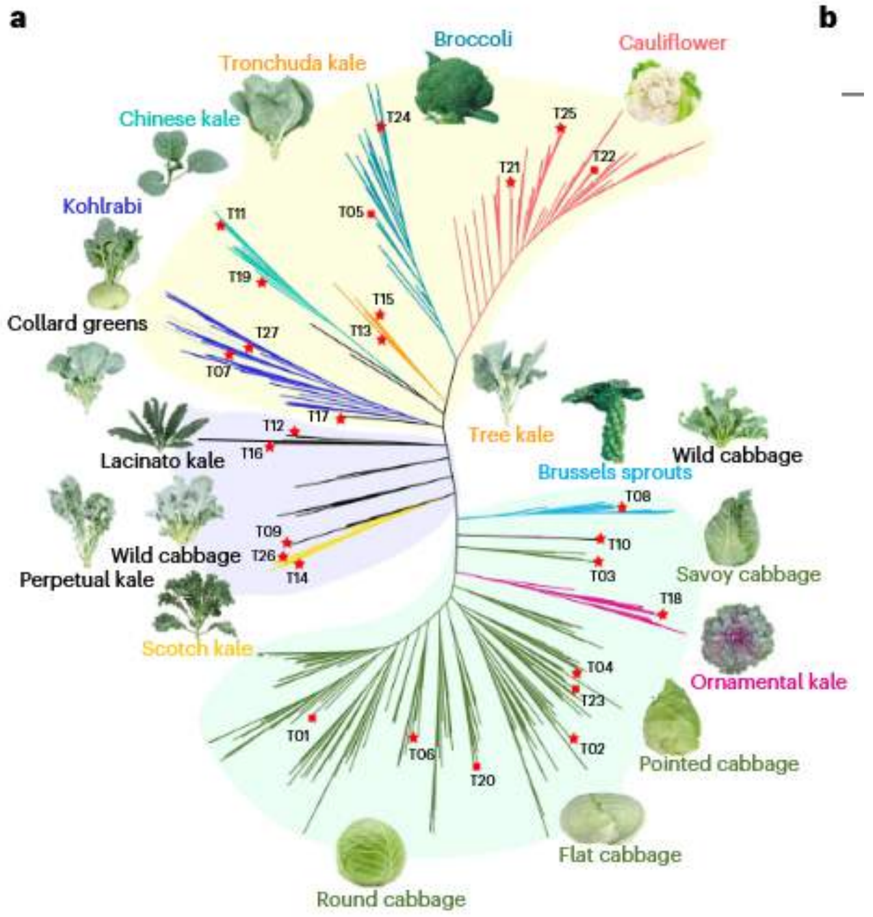
Reference Genomes

- Traditional methods involve mapping sample reads to reference
- Reference is assumed to be ground truth for orienting reads
- True rearrangements, duplications, deletions lost or thrown out during QC



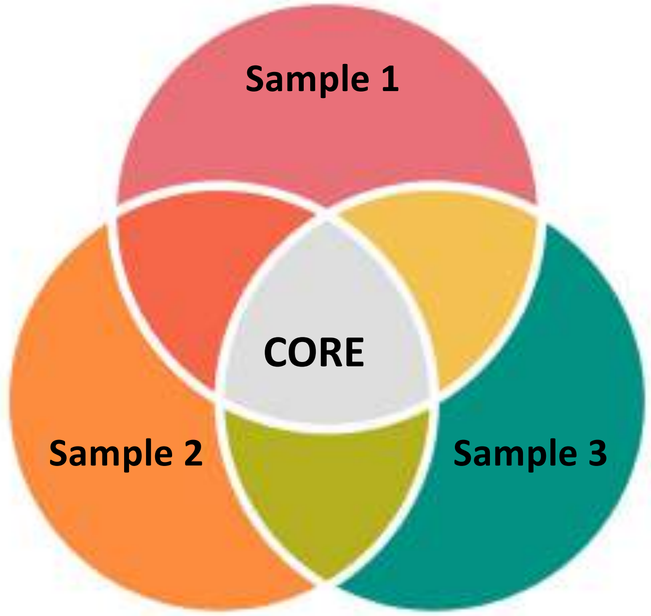
Reference Genomes

- These rearrangements… aka Structural variants (SVs) tied to commercial traits in other agriculture species (Li *et al*., 2024, *nature genetics*)



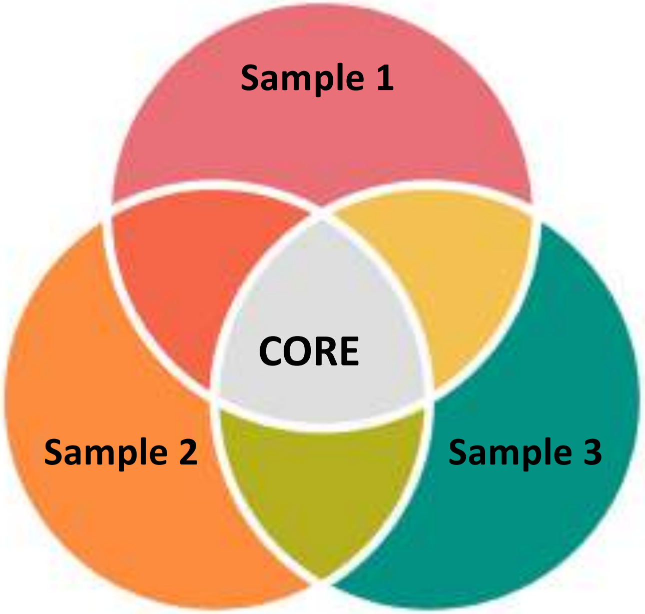
Pangenomics

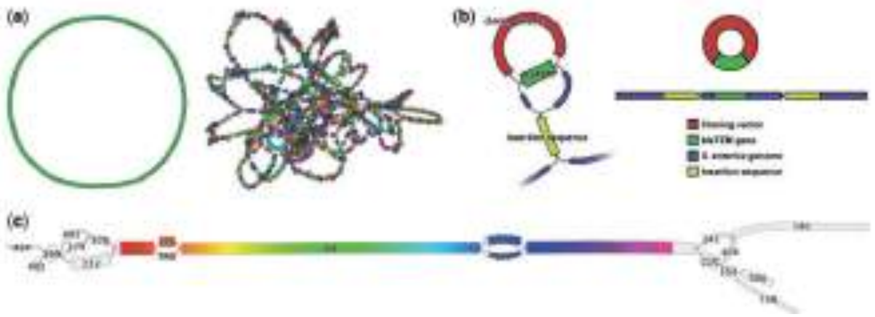
- Incorporates multiple assemblies into a pangenome to better characterize population diversity



Pangenomics

- Incorporates multiple assemblies into a pangenome to better characterize population diversity
- Many describe this diversity in graphical format; connections and paths that better represent the dynamic nature of population genetics

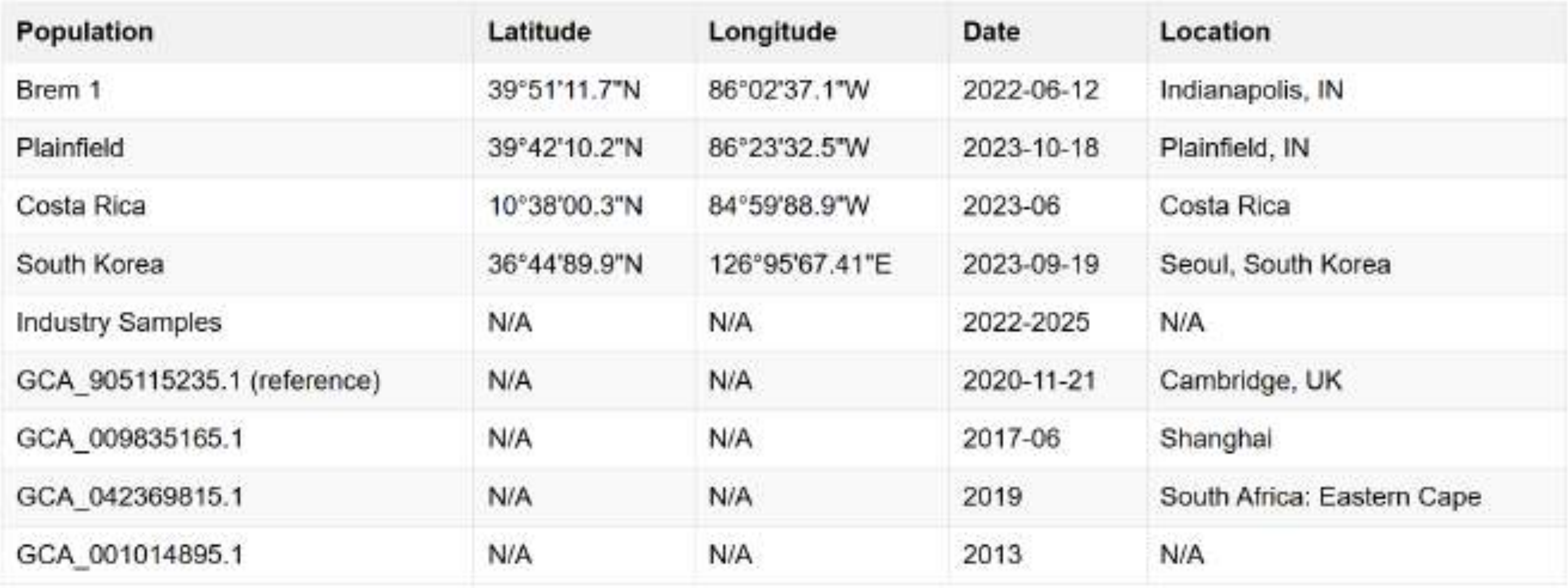




Wick *et al.*, 2015, *Bioinformatics*

Methods: data acquisition

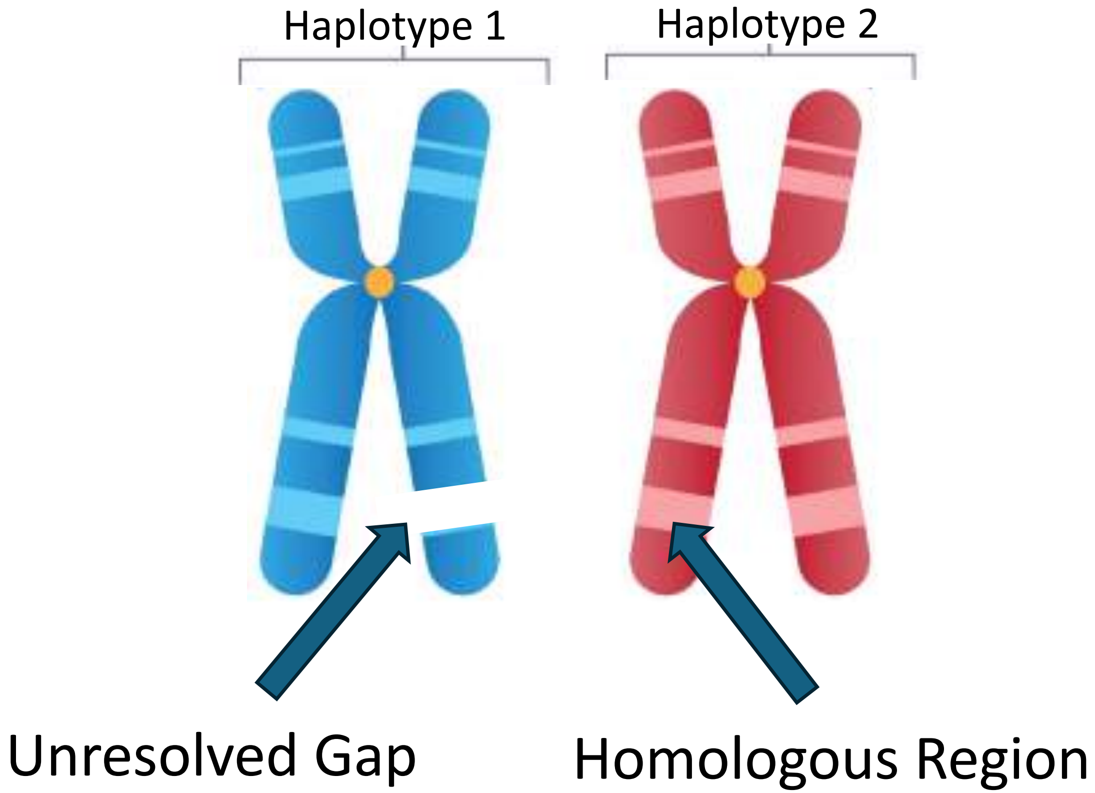
- NCBI assemblies(datasets-cli)
- Reference guided and *de novo* genomes
- Wild and Domestication samples.



Population	Latitude	Longitude	Date	Location
Brem 1	39°51'11.7"N	86°02'37.1"W	2022-06-12	Indianapolis, IN
Plainfield	39°42'10.2"N	86°23'32.5"W	2023-10-18	Plainfield, IN
Costa Rica	10°38'00.3"N	84°59'88.9"W	2023-06	Costa Rica
South Korea	36°44'89.9"N	126°95'67.41"E	2023-09-19	Seoul, South Korea
Industry Samples	N/A	N/A	2022-2025	N/A
GCA_905115235.1 (reference)	N/A	N/A	2020-11-21	Cambridge, UK
GCA_009835165.1	N/A	N/A	2017-06	Shanghai
GCA_042369815.1	N/A	N/A	2019	South Africa: Eastern Cape
GCA_001014895.1	N/A	N/A	2013	N/A

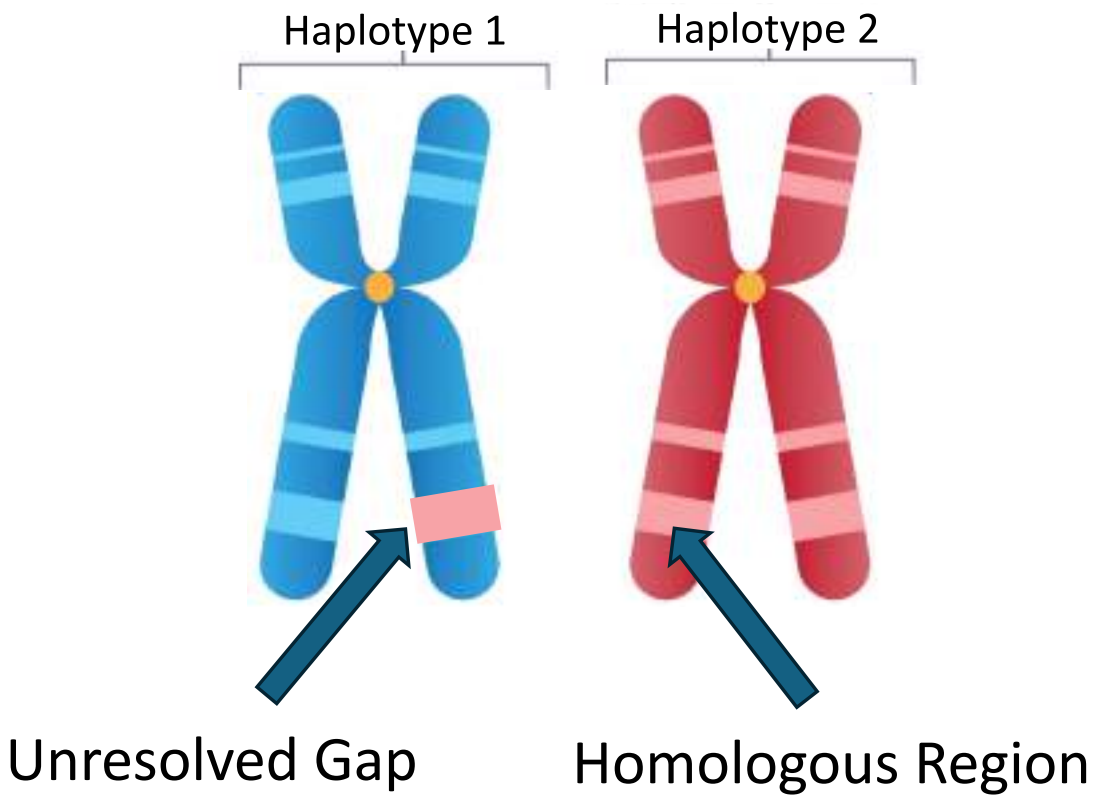
Methods *De novo* BSF assembly

- Hifi long reads assembled with *Hifiasm (*Cheng *et al*., 2021, *Nat Methods*)



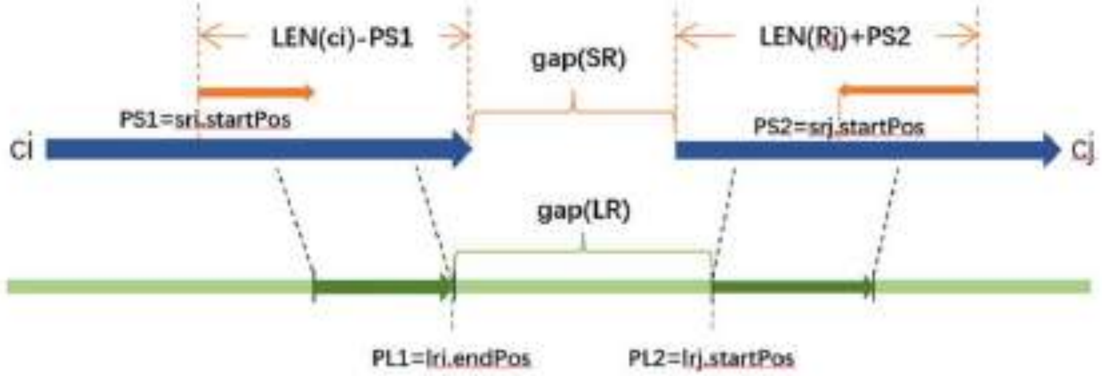
Methods *De novo* BSF assembly

- Hifi long reads assembled with *Hifiasm* (Cheng *et al*., 2021, *Nat Methods*)
- split into haplotypes and self scaffolded



Methods *De novo* BSF assembly

- Hifi long reads assembled with Hifiasm *(*Cheng *et al*., 2021, *Nat Methods*)
- split into haplotypes and self scaffolded
- Hybrid scaffolder SLHSD, Nanopore + Illumina reads alone, no Hi-C/Omni-C (Luo *et al.,* 2023, *Briefings in Bioinformatics*)

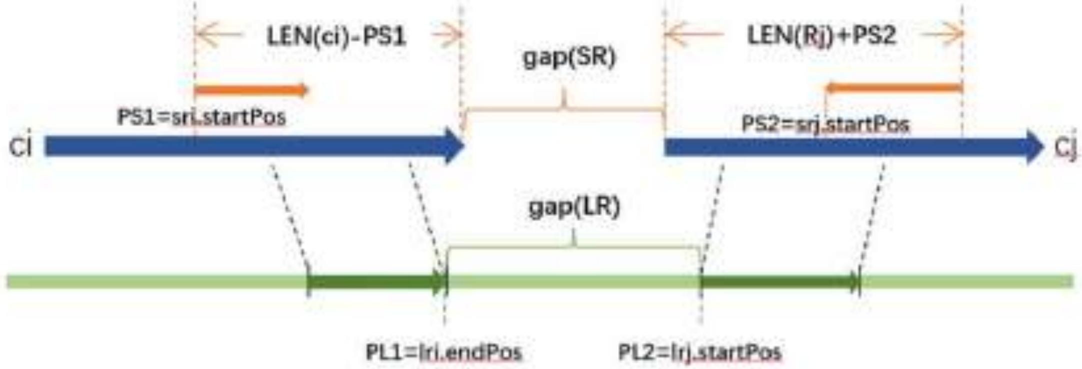


Luo *et al*., 2023, *Briefings in Bioinformatics*

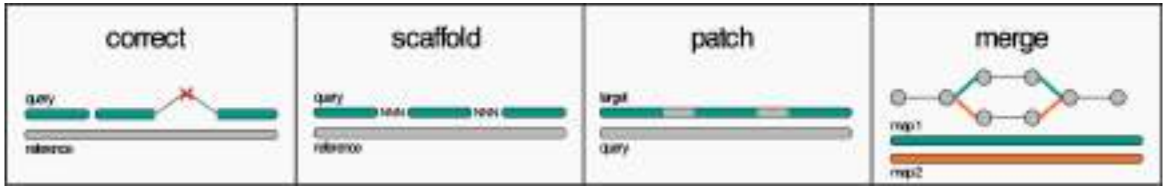
Methods *De novo** BSF assembly

- Hifi long reads assembled with Hifiasm
- split into haplotypes and self scaffolded
- Hybrid scaffolder SLHSD, Nanopore + Illumina reads alone, no Hi-C/Omni-C (Luo *et al.,* 2023, *Briefings in Bioinformatics*)

*Ragtag to order scaffolds based on reference (Alonge *et al.,* 2022, *Genome Biology*)



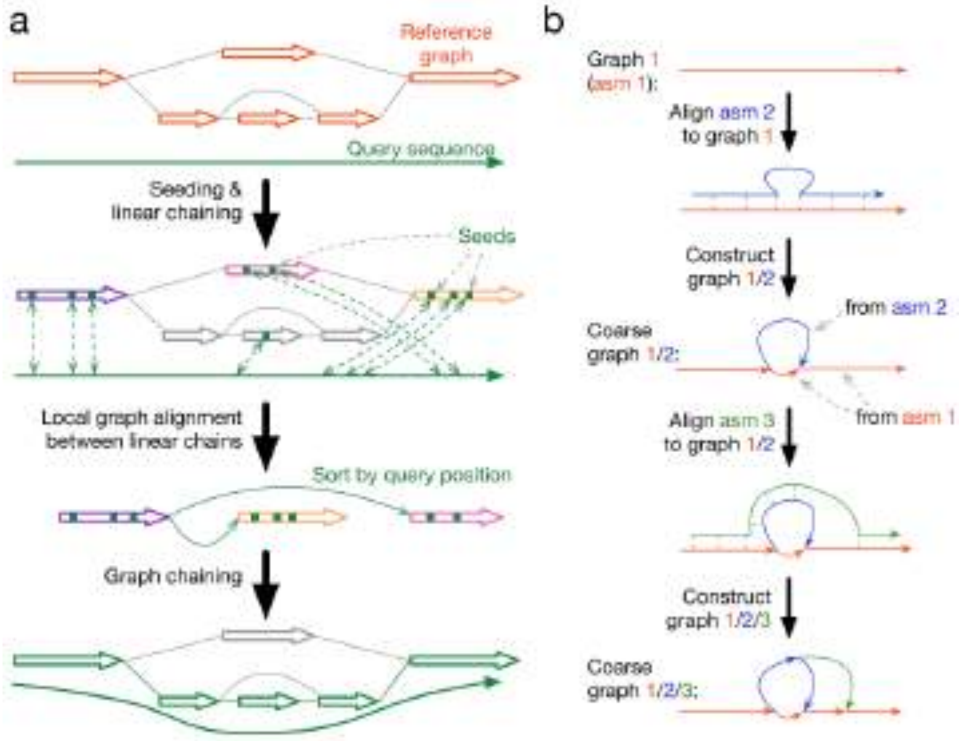
Luo *et al.*, 2023, *Briefings in Bioinformatics*



Alonge *et al.*, 2022, *Genome Biology*

Methods: pangenome assembly/graphing

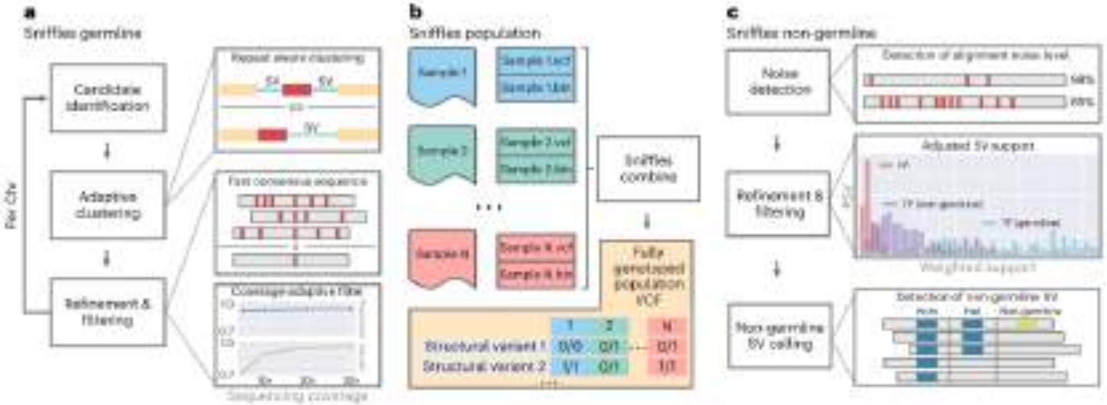
- *Minigraph* assembler (Li *et al*., 2020, *Genome Biology*)
- Quick, simple (single line of code)
- Variation is shown in graphical format: better reflects population diversity and reduces file size



Li *et al*., 2020, *Genome Biology*

Methods: TE and SV calling

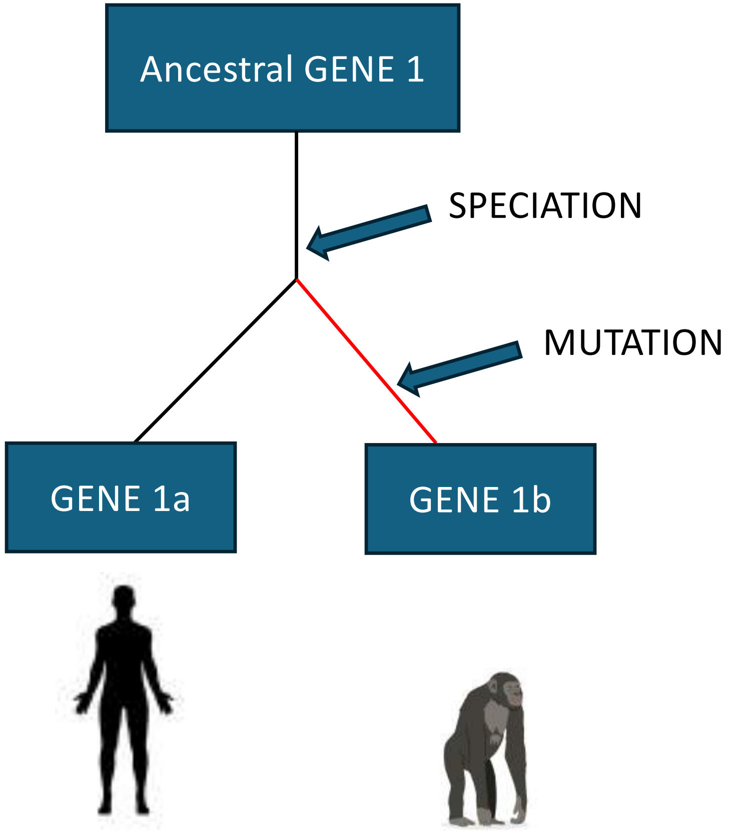
- *Sniffles* pipeline to call SVs pairwise (Smolka *et al*., 2024, *Nat Biotechnol*)



Smolka *et al*., 2024, *Nat Biotechnol*

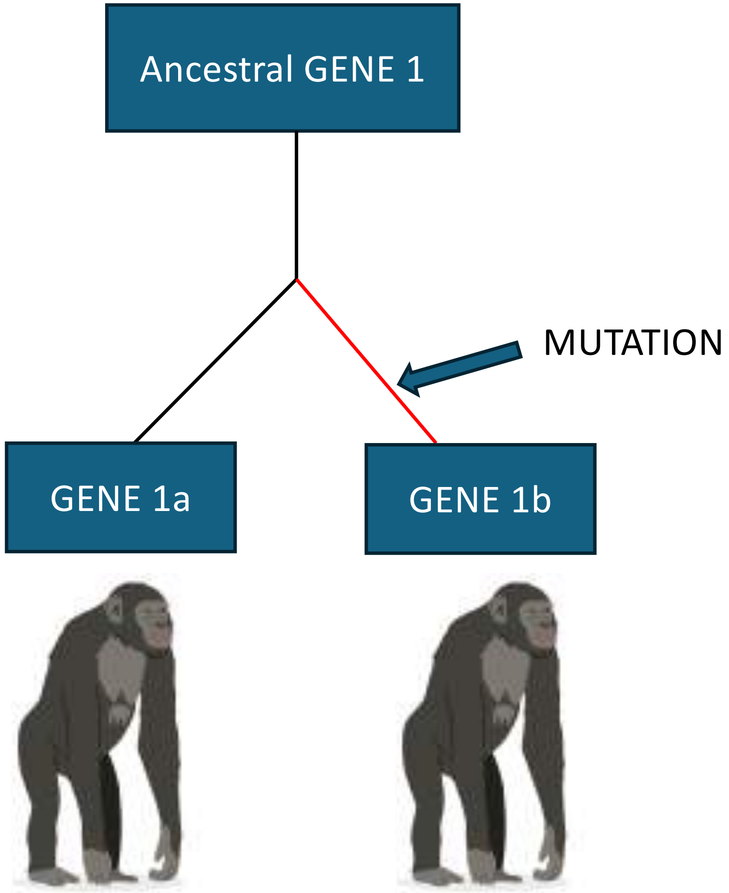
Methods: Orthfinder

- Orthology traditionally involves genes from different species, common ancestor



Methods: Orthology

- Orthology traditionally involves genes from different species, common ancestor
- Similar idea, cluster genes with similar sequences/function



Methods: Orthfinder

- Orthology traditionally involves genes from different species, common ancestor
- Similar idea, cluster genes with similar sequences/function
- Classify into four main categories, are there lots of unique genes?

🟩 CORE: All Samples

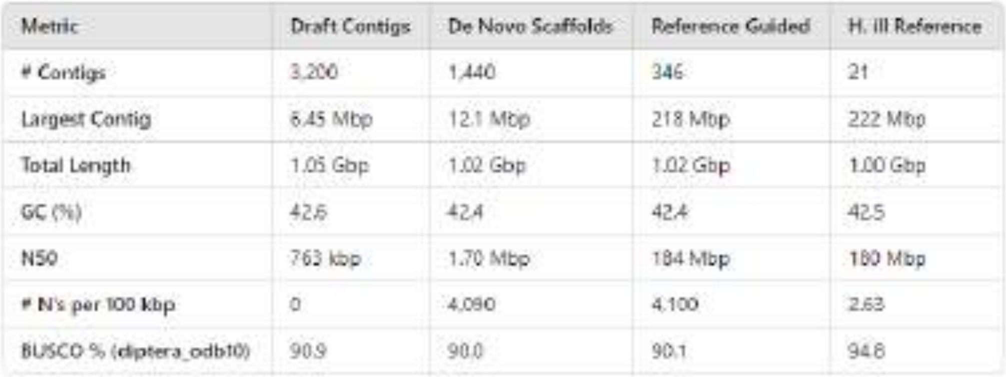
🟩 SOFTCORE: 80-99%

🟨 DISPENSABLE: 2-79%

🟥 PRIVATE: 1 Sample

Results: *De novo* assembly

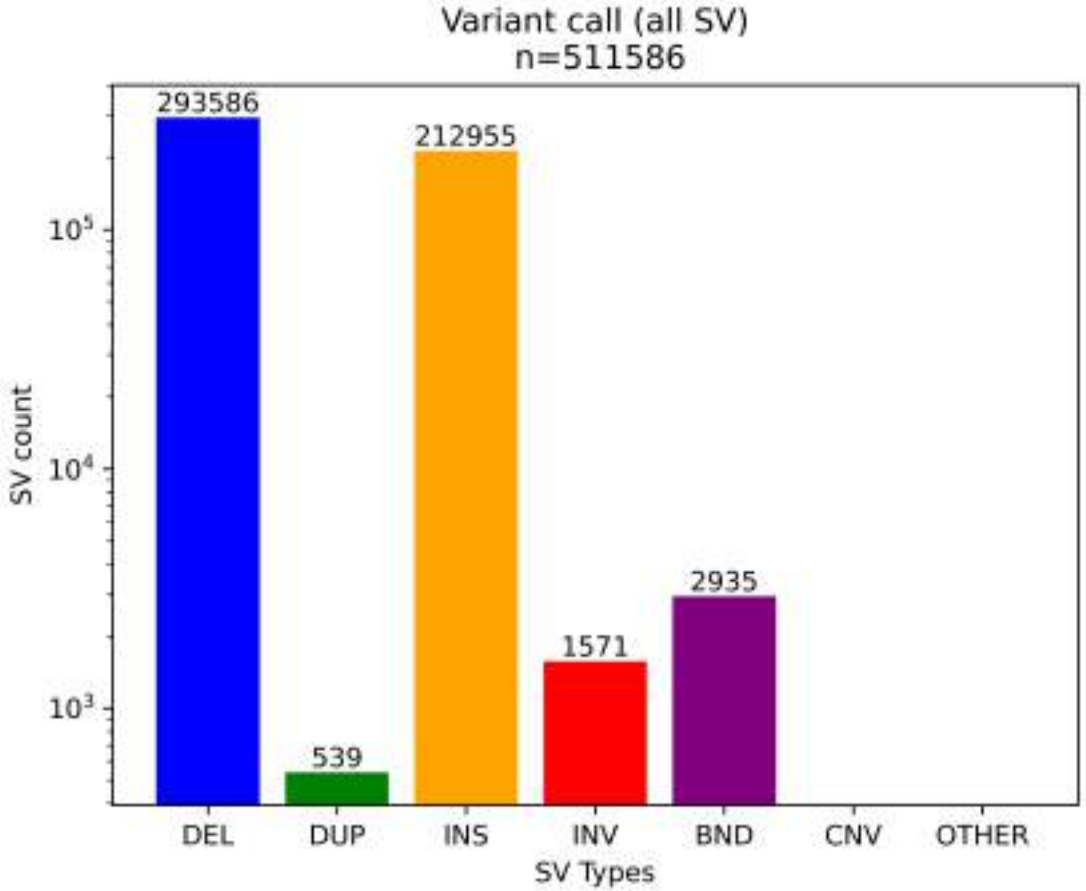
- Initial hifiasm assembly not contiguous
- Scaffolding offered modest improvements in contiguity (N50=1.7Mbp, largest contig= 12.1 Mbp)
- Scaffolds large enough to detect SVs



Metric	Draft Contigs	De Novo Scaffolds	Reference Guided	H. ill Reference
# Contigs	3,200	1,440	346	21
Largest Contig	6.45 Mbp	12.1 Mbp	218 Mbp	222 Mbp
Total Length	1.05 Gbp	1.02 Gbp	1.02 Gbp	1.00 Gbp
GC (%)	42.6	42.4	42.4	42.5
N50	763 kbp	1.70 Mbp	184 Mbp	180 Mbp
# N's per 100 kbp	0	4,090	4,100	2.63
BUSCO % (diptera_odb10)	90.9	90.0	90.1	94.8

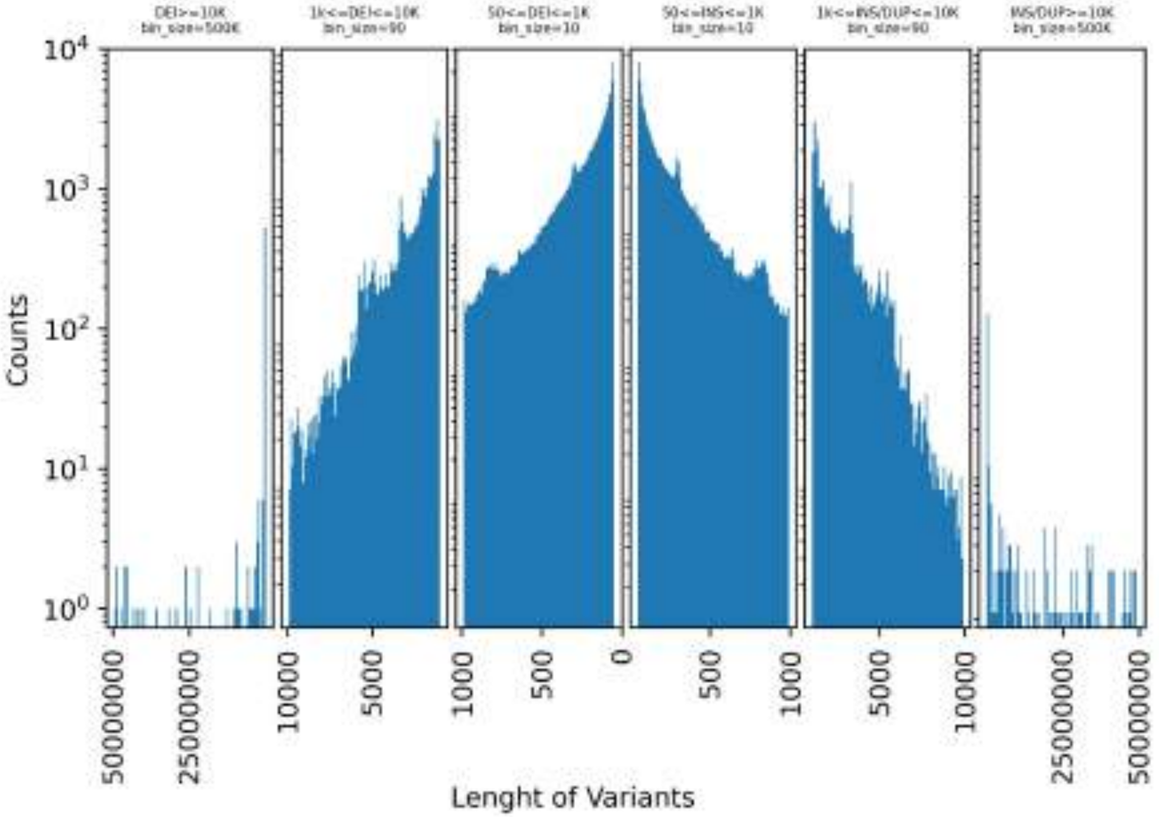
Wild *De novo* vs *H. ill* reference

- Primarily insertions/ deletions
- Small number of duplications and inversions detected



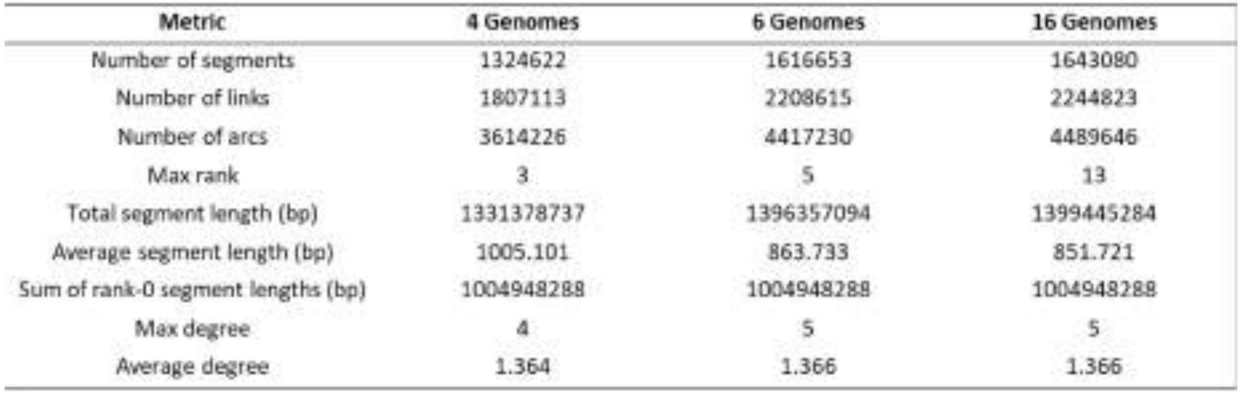
Wild *De novo* vs *H. ill* reference

- Primarily insertions/ deletions
- Small number of duplications and inversions detected
- Large inversions detected (>50Mbp)



Pangenome

- Number of segments and interconnectivity increases with additional genomes, but not linearly.

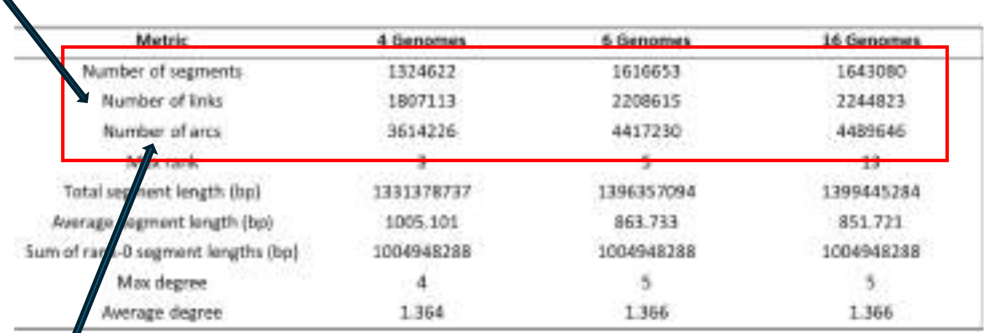


Metric	4 Genomes	6 Genomes	16 Genomes
Number of segments	1324622	1616653	1643080
Number of links	1807113	2208615	2244823
Number of arcs	3614226	4417230	4489646
Max rank	3	5	13
Total segment length (bp)	1331378737	1396357094	1399445284
Average segment length (bp)	1005.101	863.733	851.721
Sum of rank-0 segment lengths (bp)	1004948288	1004948288	1004948288
Max degree	4	5	5
Average degree	1.364	1.366	1.366

Pangenome

- Number of segments and interconnectivity increases with additional genomes, but not linearly.

More links= more connections between sections of genomes



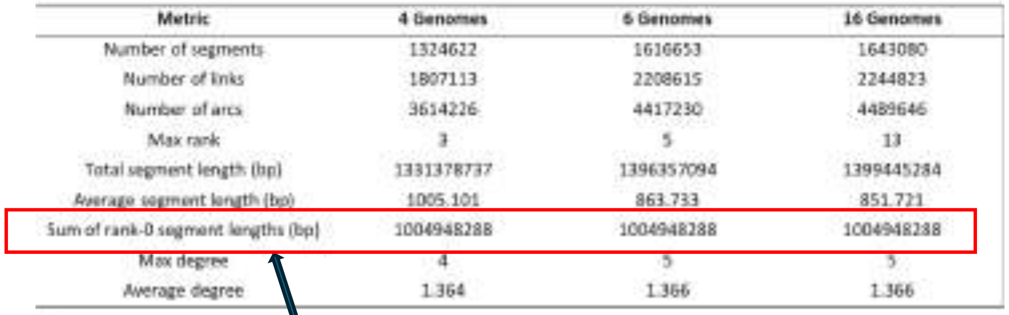
Metric	4 Genomes	6 Genomes	16 Genomes
Number of segments	1324622	1616653	1643080
Number of links	1807113	2208615	2244823
Number of arcs	3614226	4417230	4489646
Max rank	3	5	13
Total segment length (bp)	1331378737	1396357094	1399445284
Average segment length (bp)	1005.101	863.733	851.721
Sum of rank-0 segment lengths (bp)	1004948288	1004948288	1004948288
Max degree	4	5	5
Average degree	1.364	1.366	1.366

Table 3. Summary statistics (GFATools stat v0.4-r214-dirty) for pangenome assembly (Minigraph v0.21) generated by inputs from wild and domesticated black soldier fly genomes.

More arcs= more unique paths

Pangenome

- Number of segments and interconnectivity increases with additional genomes, but not linearly.
- Length shared between all genomes (rank-o segments) does not change.

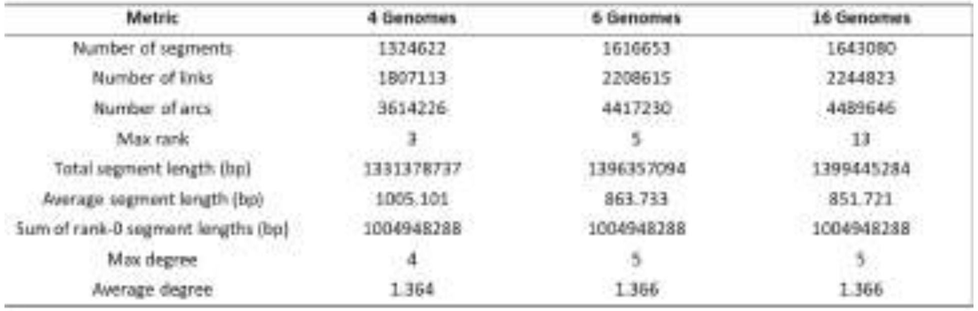


Metric	4 Genomes	6 Genomes	16 Genomes
Number of segments	1324622	1616653	1643080
Number of links	1807113	2208615	2244823
Number of arcs	3614226	4417230	4489646
Max rank	3	5	13
Total segment length (bp)	1331378737	1396357094	1399445284
Average segment length (bp)	1005.101	863.733	851.721
Sum of rank-0 segment lengths (bp)	1004948288	1004948288	1004948288
Max degree	4	5	5
Average degree	1.364	1.366	1.366

Sum rank 0: total # of base pairs shared by all input genome

Pangenome

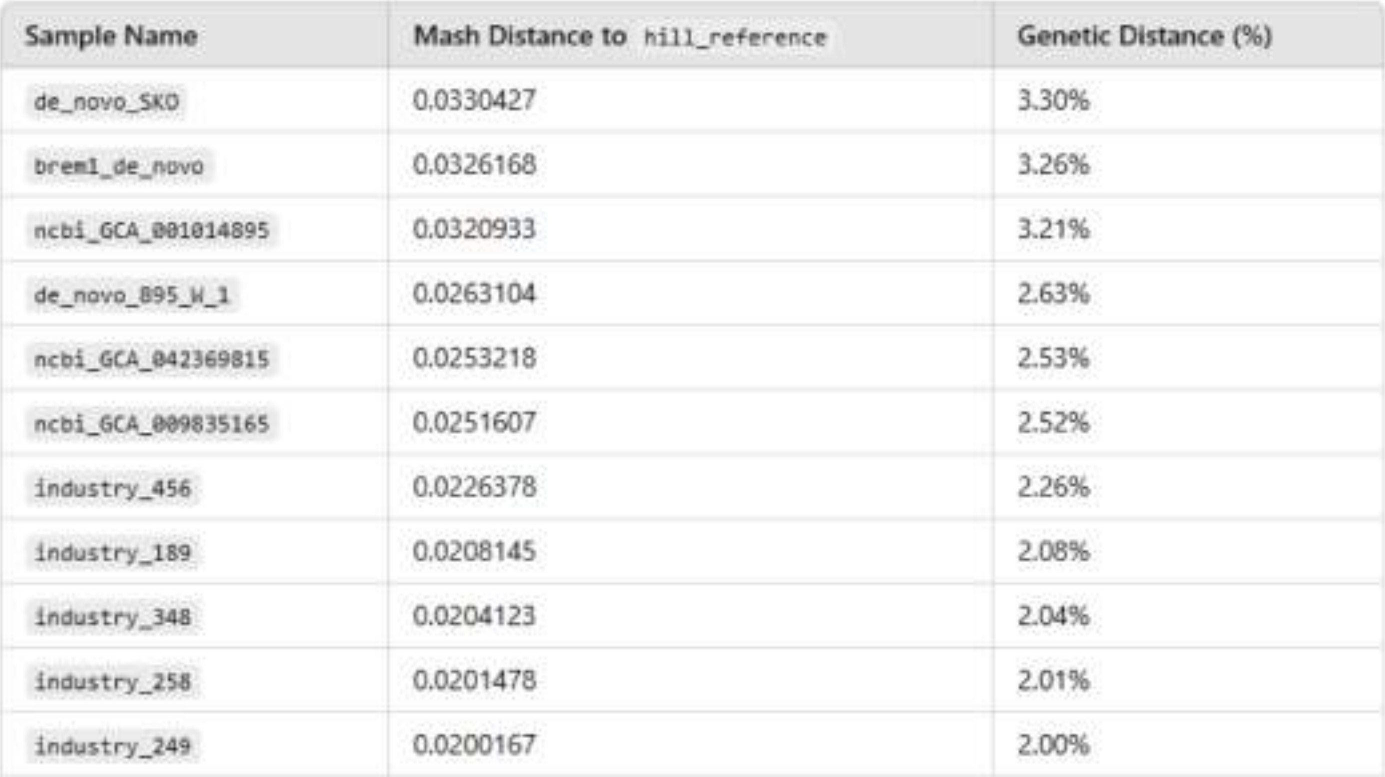
- Adding wild samples adds diversity and complexity, pointing to possible rearrangements and large INDELs
- The draft genome likely captures the conserved core genome



Metric	4 Genomes	6 Genomes	16 Genomes
Number of segments	1324622	1616653	1643080
Number of links	1807113	2208615	2244823
Number of arcs	3614226	4417230	4489646
Max rank	3	5	13
Total segment length (bp)	1331378737	1396357094	1399445284
Average segment length (bp)	1005.101	863.733	851.721
Sum of rank-0 segment lengths (bp)	1004948288	1004948288	1004948288
Max degree	4	5	5
Average degree	1.364	1.366	1.366

Pangenome

- Large whole-genome genetic divergence (>3%) between samples
- Not more than observed in papers based on COI (~5%, Stahls *et al.*, 2020, *BMC Evol Biol.)*

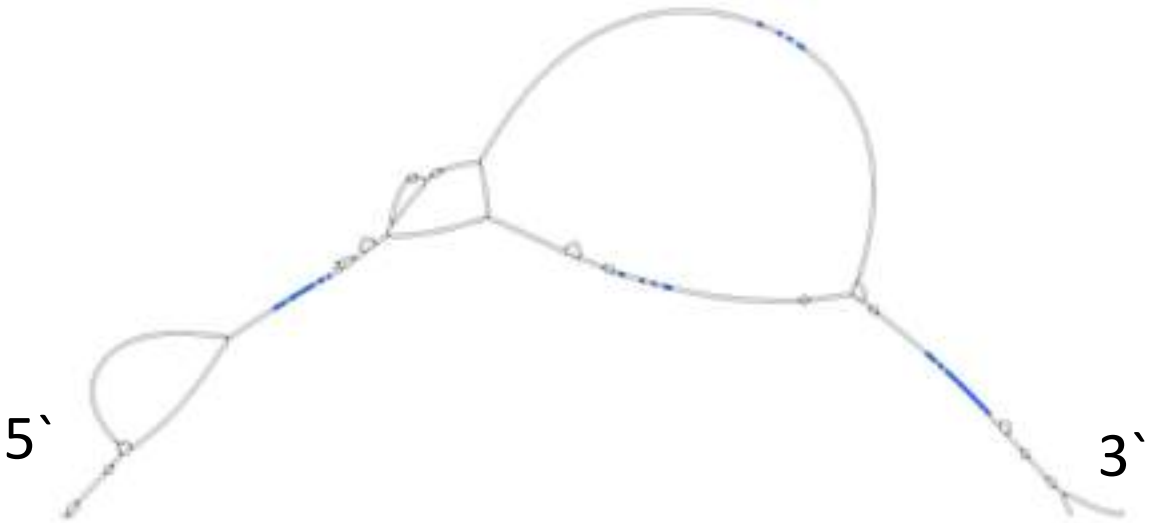


Sample Name	Mash Distance to `hill_reference`	Genetic Distance (%)
`de_novo_SKO`	0.0330427	3.30%
`brem1_de_novo`	0.0326168	3.26%
`ncbi_GCA_001014895`	0.0320933	3.21%
`de_novo_895_W_1`	0.0263104	2.63%
`ncbi_GCA_042369815`	0.0253218	2.53%
`ncbi_GCA_009835165`	0.0251607	2.52%
`industry_456`	0.0226378	2.26%
`industry_189`	0.0208145	2.08%
`industry_348`	0.0204123	2.04%
`industry_258`	0.0201478	2.01%
`industry_249`	0.0200167	2.00%

Pangenome

- Graphical approach allows for visualization of genes of interest
- BSF *InR* mostly conserved among samples, but split paths show variability in region matching Furin-like cysteine rich domain

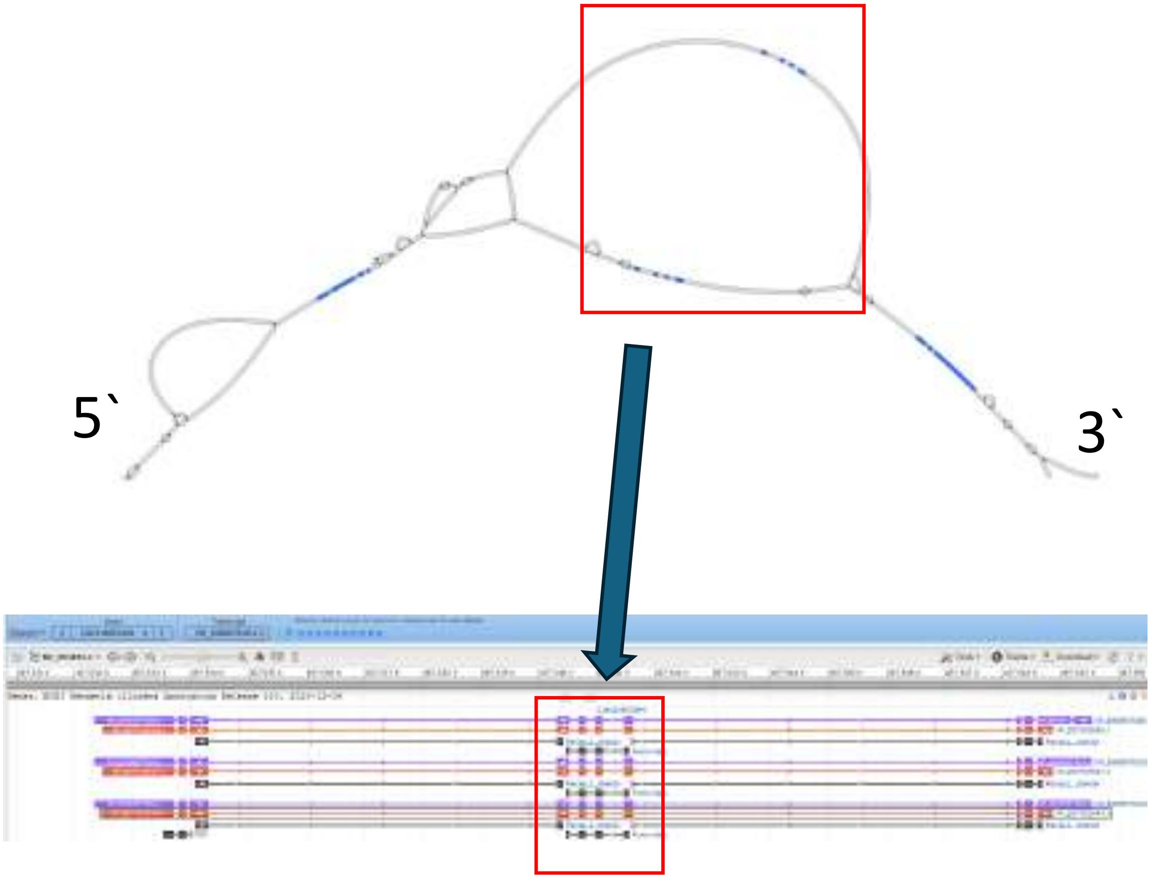
Black soldier fly *Insulin-like Receptor*



Pangenome

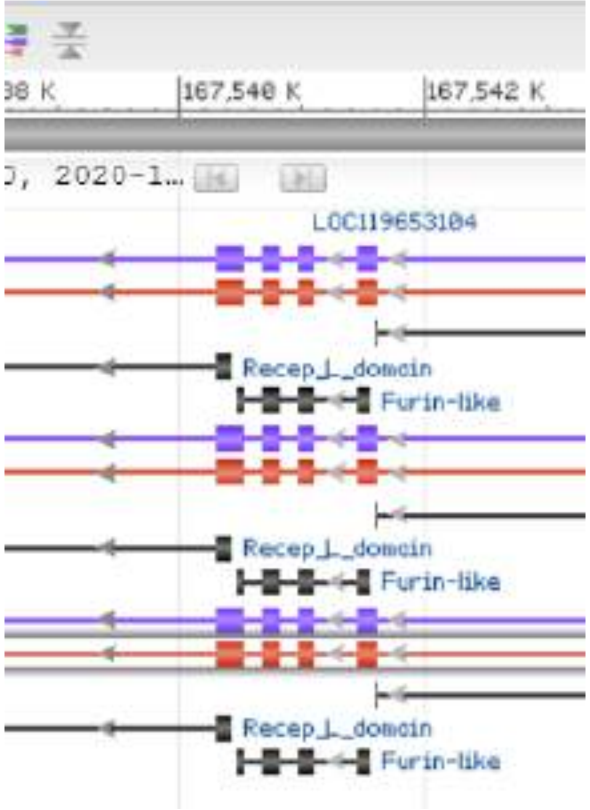
- Graphical approach allows for visualization of genes of interest
- BSF *InR* mostly conserved among samples, but split paths show variability in region matching Furin-like cysteine rich domain

Black soldier fly *Insulin-like Receptor*



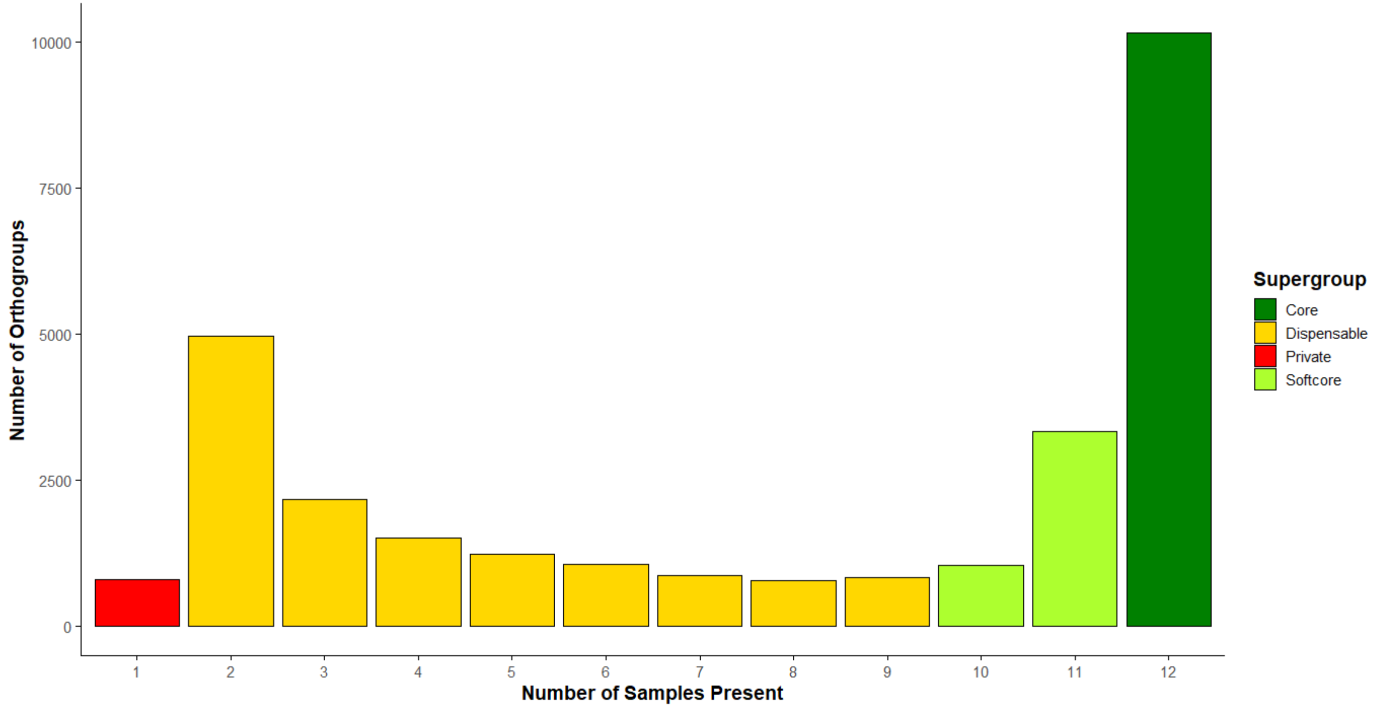
Pangenome

- Graphical approach allows for visualization of genes of interest
- BSF *InR* mostly conserved among samples, but split paths show variability in region matching Furin-like cysteine rich domain



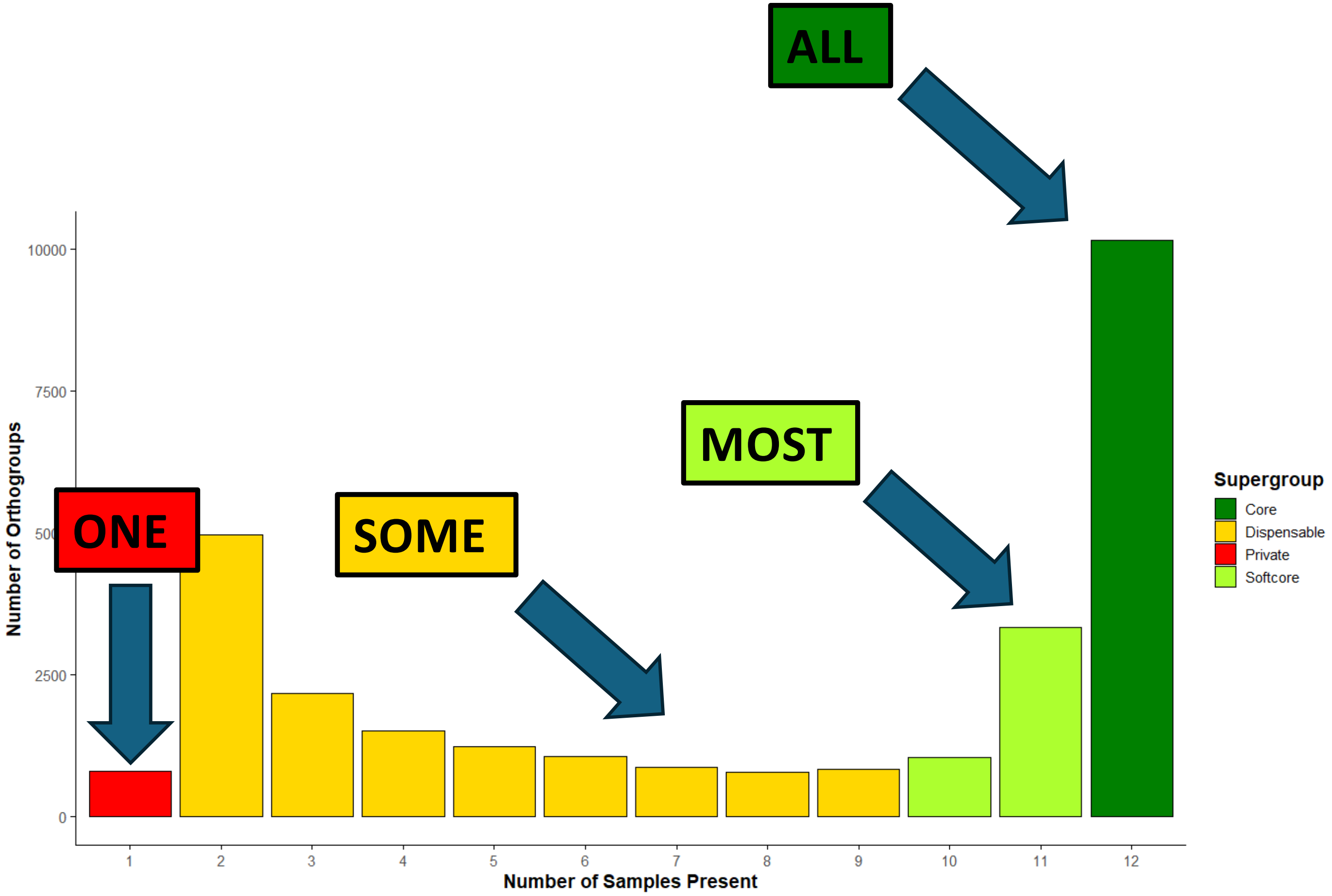
BSF Diversity

- *Hermetia illucens* orthology analysis reveals diversity even in this small set.



BSF Diversity

- *Hermetia illucens* orthology analysis reveals diversity even in this small set.



Conclusions

- Draft pangenome captures much of the core Black Soldier Fly genome
- Incorporates newly described diversity from wild samples across South Korea, Central America, and the United States (geographically close to Sheppard strain origin)
- Room to grow: more samples, pangenome aware analysis tools (GraffiTE, Odgi, Cactus)

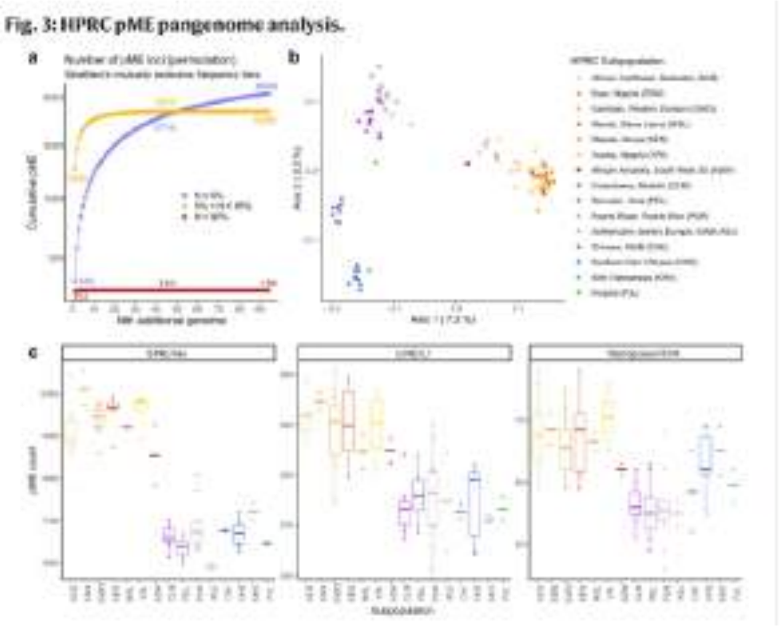


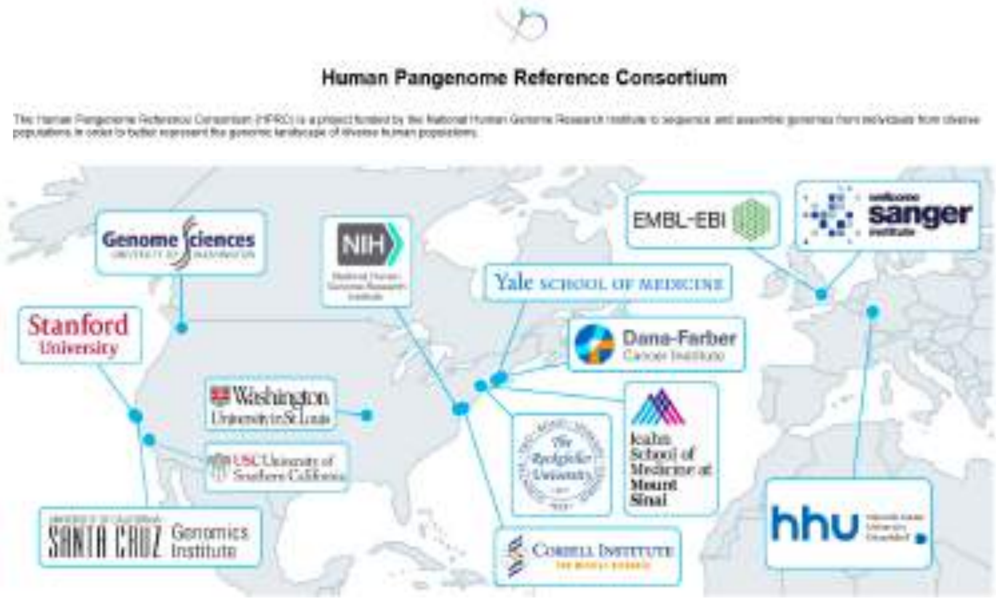
Fig. 3: HPRC pME pangenome analysis.

Groza et al., 2024, *Nat Commun*

Conclusions

https://humanpangenome.org/

- The pangenome is a group effort, incorporating existing assemblies, with potential for future researchers to contribute their own.



Human Pangenome Reference Consortium

The Human Pangenome Reference Consortium (HPRC) is a project funded by the National Human Genome Research Institute to sequence and assemble genomes from individuals from diverse populations in order to better represent the genomic landscape of diverse human populations.

Genome Sciences University of Washington
NIH
EMBL-EBI
wellcome sanger institute
Yale SCHOOL OF MEDICINE
Stanford University
Dana-Farber Cancer Institute
Washington University in St. Louis
USC University of Southern California
The Rockefeller University
Icahn School of Medicine at Mount Sinai
University of California Santa Cruz Genomics Institute
Coriell Institute
hhu

Conclusions

The pangenome is a group effort, incorporating existing assemblies, with potential for future researchers to contribute their own.

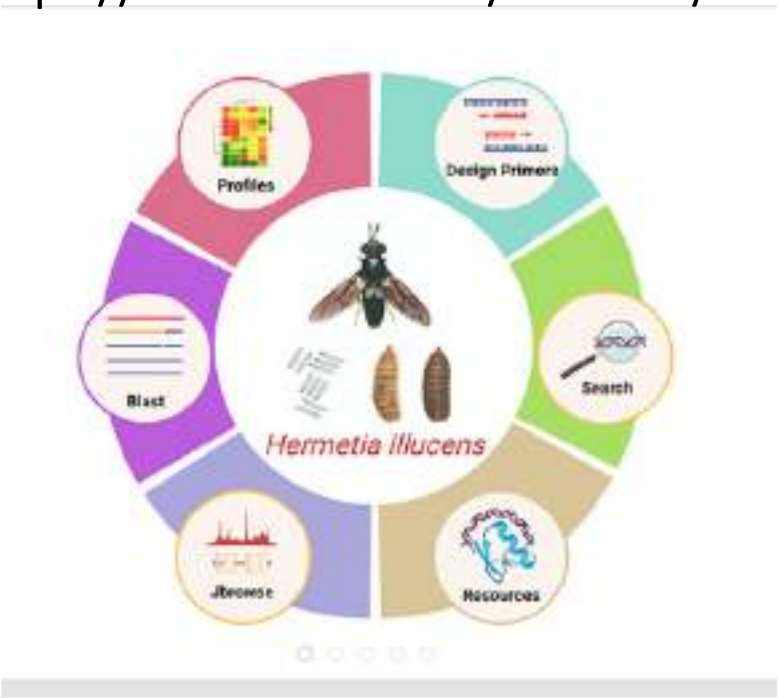
Trend towards centralization analogous to Flybase, tools like BSFbase are just a start (Dong et al., 2023, Insect Science).

https://insectomics.net/BSFbase/

Conclusions

- The pangenome is a group effort, incorporating existing assemblies, with potential for future researchers to contribute their own.
- Trend towards centralization analogous to Flybase, tools like BSFbase are just a start (Dong *et al.,* 2023, *Insect Science*).
- Insect agriculture benefits from collaboration

https://insectomics.net/BSFbase/



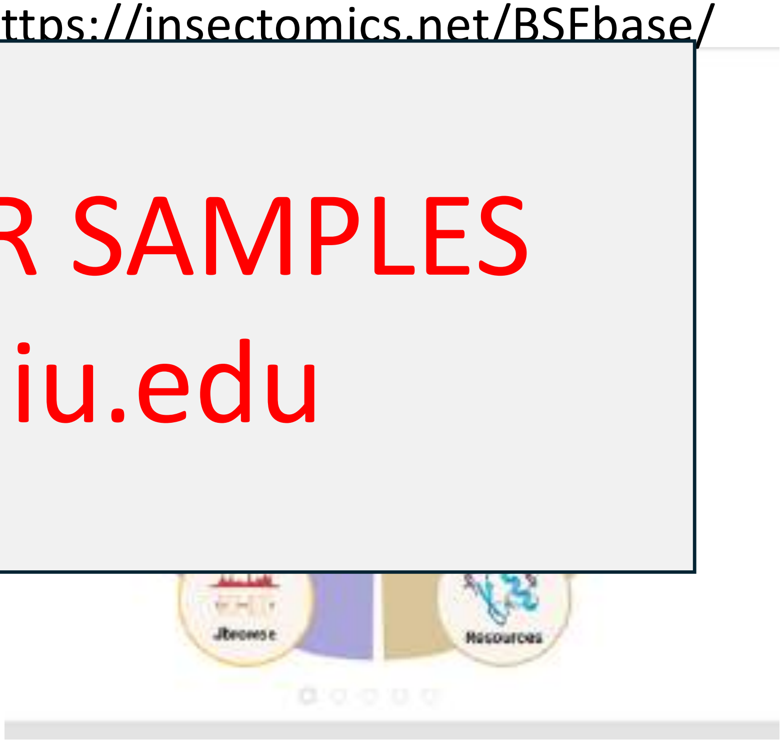
Conclusions

- The
  effo
  asse
  futu
  thei
- Tren
  anal
  BSFbase are just a start (Dong *et al.*, 2023, *Insect Science*).

SEND US YOUR SAMPLES

hrosche@iu.edu

https://insectomics.net/BSFbase/



Acknowledgments

- This work was supported by the National Science Foundation through the Industry-University Cooperative Research Centers (IUCRC) NSF cooperative agreements NSF-IIP-2052454 (TAMU), NSF-IIP-2052565 (IUI), & NSF-IIP-2052788 (MSU).

- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Industry Advisory Board Members of the Center for Environmental Sustainability through Insect Farming.

Questions?