Transposons: a blessing curse

Scott Teresi

Department of Horticulture
Genetics & Genome Sciences Program
IMPACTS

Monday 28th October, 2024

Table of Contents:

Personal Introduction

- B.S in Biology from The College of William & Mary (2019)
  - Minor in Computational & Applied Math & Statistics

Personal Introduction

- B.S in Biology from The College of William & Mary (2019)
  - Minor in Computational & Applied Math & Statistics
- Joined the Puzey Lab
  - Was able to collaborate with Pat's lab on the strawberry genome project

Article | Open access | Published: 25 February 2019

Origin and evolution of the octoploid strawberry genome

Patrick P. Edger, Thomas J. Poorten, Robert VanBuren, Michael A. Hardigan, Marivi Colle, Michael R. McKain, Ronald D. Smith, Scott J. Teresi, Andrew D. L. Nelson, Ching Man Wai, Elizabeth I. Alger, Kevin A. Bird, Alan E. Yocca, Nathan Pumplin, Shujun Ou, Gil Ben-Zvi, Avital Brodt, Kobi Baruch, Thomas Swale, Lily Shiue, Charlotte B. Acharya, Glenn S. Cole, Jeffrey P. Mower, Kevin L. Childs, ... Steven J. Knapp

+ Show authors

Nature Genetics **51**, 541–547 (2019) | Cite this article

Personal Introduction

- B.S in Biology from The College of William & Mary (2019)
  - Minor in Computational & Applied Math & Statistics
- Joined the Puzey Lab
  - Was able to collaborate with Pat's lab on the strawberry genome project
- Joined Pat's Lab for the 2018 Plant Genomics REU
  - More on this later!

Personal Introduction

- B.S in Biology from The College of William & Mary (2019)
  - Minor in Computational & Applied Math & Statistics
- Joined the Puzey Lab
  - Was able to collaborate with Pat's lab on the strawberry genome project
- Joined Pat's Lab for the 2018 Plant Genomics REU
  - More on this later!
- Returned to Pat's Lab for my PhD in Fall 2019

What are TEs?

- Transposable elements (TEs, transposons) are mobile, repetitive, genetic entities.

What are TEs?

- Transposable elements (TEs, transposons) are mobile, repetitive, genetic entities.
- TEs are ubiquitous components of most genomes, and can often make up a large portion of the genome.

What are TEs?

- Transposable elements (TEs, transposons) are mobile, repetitive, genetic entities.
- TEs are ubiquitous components of most genomes, and can often make up a large portion of the genome.
- TEs are a major force of mutation and genome evolution.

TEs were discovered by Barbara McClintock

- Barbara McClintock discovered TEs in maize in the 1940s.

 [1]

Feschotte (2023) - Nature Reviews Genetics

[1] Image: Wikipedia

TEs were discovered by Barbara McClintock

- Barbara McClintock discovered TEs in maize in the 1940s.
- She observed that a certain region of DNA was consistently breaking and moving, resulting in kernel color changes.

 [2]

Feschotte (2023) - Nature Reviews Genetics

[2] Image: Nature Journal

TEs were discovered by Barbara McClintock

- Barbara McClintock discovered TEs in maize in the 1940s.
- She observed that a certain region of DNA was consistently breaking and moving, resulting in kernel color changes.
- She called these oddities “controlling elements”, due to their ability to create phenotypic variation.

 [2]

Feschotte (2023) - Nature Reviews Genetics

[2] Image: Nature Journal

TEs were discovered by Barbara McClintock

- Barbara McClintock discovered TEs in maize in the 1940s.
- She observed that a certain region of DNA was consistently breaking and moving, resulting in kernel color changes.
- She called these oddities "controlling elements", due to their ability to create phenotypic variation.
- However, her work was largely met with skepticism because it was so far ahead of its time and limited to maize.

 [2]

Feschotte (2023) - Nature Reviews Genetics

[2] Image: Nature Journal

So why should we care about TEs?

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer

Review Article | Published: 09 June 2017

Transposable elements in cancer

Kathleen H. Burns

Nature Reviews Cancer **17**, 415–424 (2017) | Cite this article

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer
- TEs contribute to disease resistance in a variety of organisms

RESEARCH Open Access

CrossMark

New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication

Seungill Kim[1], Jieun Park[1,2], Seon-In Yeom[3], Yong-Min Kim[4], Eunyoung Seo[1], Ki-Tae Kim[5], Myung-Shin Kim[1], Je Min Lee[6], Kyeongchae Cheong[2,5], Ho-Sub Shin[1], Saet-Byul Kim[1], Koeun Han[1,7], Jundae Lee[8], Minkyu Park[9], Hyun-Ah Lee[1], Hye-Young Lee[1], Youngsill Lee[1], Soohyun Oh[1], Joo Hyun Lee[1], Eunhye Choi[1], Eunbi Choi[1], So Eui Lee[1], Jongbum Jeon[2], Hyunbin Kim[2], Gobong Choi[2], Hyeunjeong Song[2], JunKi Lee[1], Sang-Choon Lee[1], Jin-Kyung Kwon[1,7], Hea-Young Lee[1,7], Namjin Koo[4], Yunji Hong[4], Ryan W. Kim[4], Won-Hee Kang[3], Jin Hoe Huh[1], Byoung-Cheorl Kang[1,7], Tae-Jin Yang[1], Yong-Hwan Lee[2,5], Jeffrey L. Bennetzen[9] and Doil Choi[1*]

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer
- TEs contribute to disease resistance in a variety of organisms
- TEs are associated with fruit color variation
  - Apples, grapes, peppers, blood oranges, strawberry and more

ARTICLE

https://doi.org/10.1038/s41467-019-09518-x OPEN

A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour

Liyi Zhang[1], Jiang Hu[2], Xiaolei Han[1], Jingjing Li[2], Yuan Gao[1], Christopher M. Richards[3], Caixia Zhang[1], Yi Tian[1], Guiming Liu[4], Hera Gul[1], Dajiang Wang[1], Yu Tian[2], Chuanxin Yang[2], Minghui Meng[2], Gaopeng Yuan[1], Guodong Kang[1], Yonglong Wu[1], Kun Wang[1], Hengtao Zhang[5], Depeng Wang[2] & Peihua Cong[1]

The Plant Cell, Vol. 24: 1242–1255, March 2012, www.plantcell.org

Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges W OA

Eugenio Butelli,[a] Concetta Licciardello,[b] Yang Zhang,[a] Jianjun Liu,[c] Steve Mackay,[a] Paul Bailey,[a] Giuseppe Reforgiato-Recupero,[b] and Cathie Martin[a,1]

[a] John Innes Centre, Norwich NR4 7UH, United Kingdom
[b] Centro di Ricerca per l'Agrumicoltura e le Colture Mediterranee, 95024 Acireale, Italy
[c] Sichuan Academy of Agricultural Sciences, Chengdu City, Sichuan 610066, China

Retrotransposon-Induced Mutations in Grape Skin Color

Shozo Kobayashi,[1]* Nami Goto-Yamamoto,[2] Hirohiko Hirochika[3]

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer
- TEs contribute to disease resistance in a variety of organisms
- TEs are associated with fruit color variation
  - Apples, grapes, peppers, blood oranges, strawberry and more
- TEs created the peppered moth phenotype

LETTER

doi:10.1038/nature17951

The industrial melanism mutation in British peppered moths is a transposable element

Arjen E. van't Hof[1]*, Pascal Campagne[1]*, Daniel J. Rigden[1], Carl J. Yung[1], Jessica Lingley[1], Michael A. Quail[2], Neil Hall[1], Alistair C. Darby[1] & Ilik J. Saccheri[1]

1

[1] Image: BBC

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer
- TEs contribute to disease resistance in a variety of organisms
- TEs are associated with fruit color variation
  - Apples, grapes, peppers, blood oranges, strawberry and more
- TEs created the peppered moth phenotype
- Metabolite diversity and morphological variation in tomato

A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit

Han Xiao,[1] Ning Jiang,[2] Erin Schaffner,[1,3]* Eric J. Stockinger,[1] Esther van der Knaap[1]†

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-17874-2 OPEN

The impact of transposable elements on tomato diversity

Marisol Domínguez[1], Elise Dugas[1], Médine Benchouaia[2], Basile Leduque[1], José M Jiménez-Gómez[3], Vincent Colot[1✉] & Leandro Quadrana[1✉]

So why should we care about TEs?

- TEs create genetic variation and they are everywhere!
- TEs create cancer
- TEs contribute to disease resistance in a variety of organisms
- TEs are associated with fruit color variation
  - Apples, grapes, peppers, blood oranges, strawberry and more
- TEs created the peppered moth phenotype
- Metabolite diversity and morphological variation in tomato
- Maize domestication, and photoperiod sensitivity

CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize

Qin Yang[a,1], Zhi Li[a,1], Wenqiang Li[b,1], Lixia Ku[c,1], Chao Wang[a], Jianrong Ye[a], Kun Li[a], Ning Yang[b], Yipu Li[a], Tao Zhong[a], Jiansheng Li[a], Yanhui Chen[c,2], Jianbing Yan[b,2], Xiaohong Yang[a,2], and Mingliang Xu[a,2]

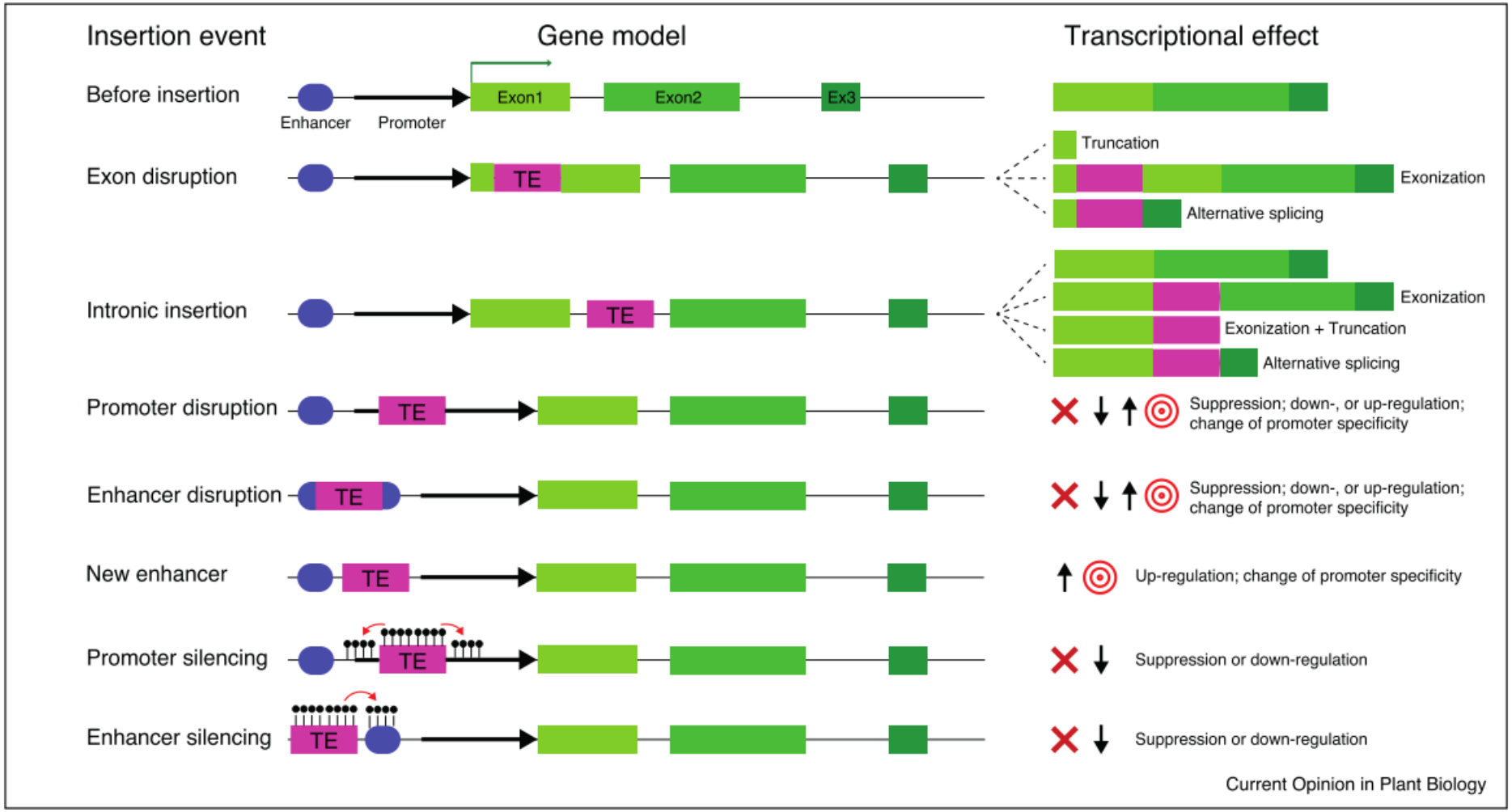
Identification of a functional transposon insertion in the maize domestication gene *tb1*

Anthony Studer[1], Qiong Zhao[1], Jeffrey Ross-Ibarra[2,3] & John Doebley[1]

So why should we care about TEs?

And many more (undiscovered) examples!

How does a TE create variation?



Dubin et al. (2018) - Current Opinion in Plant Biology

TEs have historically been maligned as “junk” DNA

- TEs were initially thought to be “junk” DNA, due to their apparent “selfish” nature. All they do is make more copies of themselves!

1

[1] Image: Forbes.com

TEs have historically been maligned as "junk" DNA

- TEs were initially thought to be "junk" DNA, due to their apparent "selfish" nature. All they do is make more copies of themselves!
- Similarities between TEs and viruses cast them as genomic invaders.

1

[1] Image: Forbes.com

TEs have historically been maligned as "junk" DNA

- TEs were initially thought to be "junk" DNA, due to their apparent "selfish" nature. All they do is make more copies of themselves!
- Similarities between TEs and viruses cast them as genomic invaders.
- Many of the new, emerging genomes in the 1990s and early 2000s such as *Homo sapiens* appeared to be filled with "dead" TEs.

[1]

[1] Image: Forbes.com

TEs have historically been maligned as “junk” DNA

- TEs were initially thought to be “junk” DNA, due to their apparent “selfish” nature. All they do is make more copies of themselves!
- Similarities between TEs and viruses cast them as genomic invaders.
- Many of the new, emerging genomes in the 1990s and early 2000s such as *Homo sapiens* appeared to be filled with “dead” TEs.
- Their high-copy number inflates genome size and sequencing costs, as well as complicates sequence alignment

1

[1] Image: Forbes.com

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status
  - Likelihood to capture, duplicate, and shuffle gene sequences

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status
  - Likelihood to capture, duplicate, and shuffle gene sequences
  - Activity — dead or alive

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status
  - Likelihood to capture, duplicate, and shuffle gene sequences
  - Activity — dead or alive
  - Intact — full-length or fragmented

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status
  - Likelihood to capture, duplicate, and shuffle gene sequences
  - Activity — dead or alive
  - Intact — full-length or fragmented
  - Copy number — low copy or abundant

TEs are difficult to study

- **There are a multitude of different TE types, each with varying characteristics and distributions!**
  - Size
  - Replication mechanism
  - Autonomous vs. non-autonomous
  - Structural features
  - Location — close to, or away from, genes
  - Insertion preference
  - Selective forces to remove
  - Methylation status
  - Likelihood to capture, duplicate, and shuffle gene sequences
  - Activity — dead or alive
  - Intact — full-length or fragmented
  - Copy number — low copy or abundant
  - TE insertions into other TEs

TEs are difficult to study

Classification		Structure	TSD	Code	Occurrence
Order	**Superfamily**				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O

Structural features

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

TEs are difficult to study

Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats — Terminal inverted repeats — Coding region — Non-coding region

Diagnostic feature in non-coding region — Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

TE Density Software — Inspiration

Our understanding of TEs is limited by the tools we have to study them.

Letter

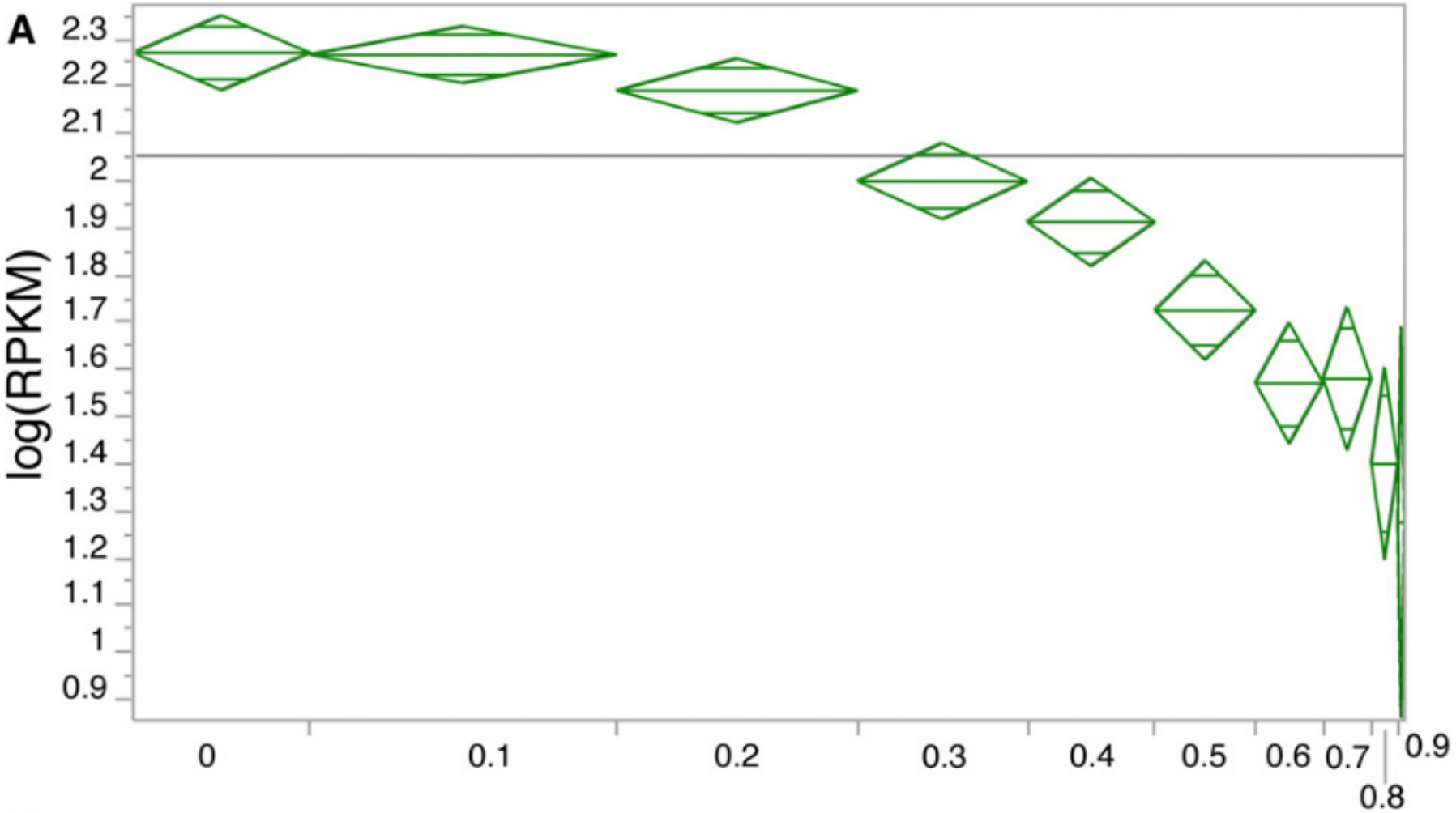
Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression

Jesse D. Hollister and Brandon S. Gaut[1]

Department of Ecology and Evolutionary Biology, University of Californina, Irvine, Irvine, California 92697-2525, USA

Hollister and Gaut (2009) - Genome Research

TE Density Software — Inspiration



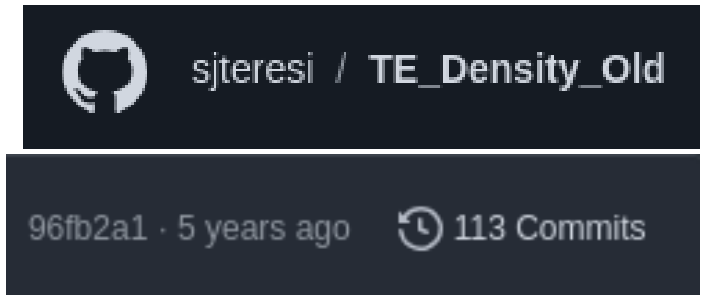
Edger et al. (2017) - Plant Cell

TE Density Software — REU Edition

- I joined Pat's lab for the summer of 2018 as part of the Plant Genomics REU.
- I wanted to develop software to calculate TE density around genes, and better understand the different associations between TEs and genes.

TE Density Software — REU Edition

- I joined Pat's lab for the summer of 2018 as part of the Plant Genomics REU.
- I wanted to develop software to calculate TE density around genes, and better understand the different associations between TEs and genes.



TE Density Software — REU Edition

I managed to get a working prototype by the end of the summer and had some cool results

KEGG Pathways			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
03010	Ribosome	30	1.5e-13
00500	Starch and sucrose metabolism	14	0.000142

PFAM Protein Domains			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
PF00982	Glycosyltransferase family 20	6	2.23e-08
PF02358	Trehalose-phosphatase	6	2.23e-08

INTERPRO Protein Domains and Features			
pathway ID	*pathway description*	*count in gene set*	*false discovery rate*
IPR001830	Glycosyl transferase, family 20	6	3.81e-08
IPR003337	Trehalose-phosphatase	6	3.81e-08
IPR006379	HAD-superfamily hydrolase, subfamily IIB	4	0.000485

pace

TE Density Software — Development

Teresi *et al. Mobile DNA* (2022) 13:11
https://doi.org/10.1186/s13100-022-00264-4

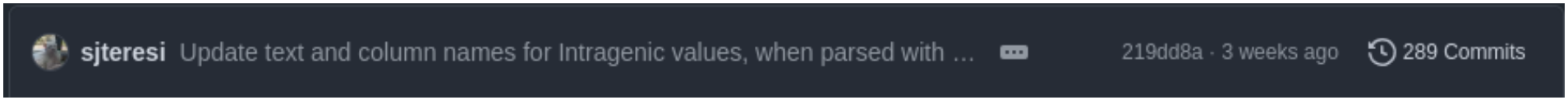
Mobile DNA

SOFTWARE Open Access

TE Density: a tool to investigate the biology of transposable elements

Scott J. Teresi[1,2], Michael B. Teresi[3] and Patrick P. Edger[1,2*]

TE Density Software — Development



- For every gene, report the amount of TE-occupied base-pairs normalized by a given search space. I.e Gene X has 50% of the 5000 BP upstream area occupied by TEs, and then that number can further be broken down into the various TE types

Algorithm 1: Calculate Density

CalculateDensity
- **inputs** : set of windows W
- **output** : density ρ
- **dataset**: pseudomolecule C
- get *genes* from C
- get *transposons* from C
- initialize overlaps O
- **foreach** *geneName* $g_j \in genes$ **do**
  - **foreach** *window* $w_j \in W$ **do**
    - // see (1) (2) (3)
    - $o = Overlap(g_j, w_j, transposons)$
    - $O[g_j, w_j] = o$
- initialize densities ρ
- **foreach** *geneName* $g_j \in genes$ **do**
  - **foreach** *TE identity* $t_j \in transposons$ **do**
    - **foreach** *window* $w_j \in W$ **do**
      - // see (4)
      - $d = Density(g_j, t_j, w_j, O_j)$
      - $\rho[g_j, w_j, t_j] = d$

TE Density Software — Development

- For every gene, report the amount of TE-occupied base-pairs normalized by a given search space. I.e Gene X has 50% of the 5000 BP upstream area occupied by TEs, and then that number can further be broken down into the various TE types
- Software really isn't that complex (just a lot of iteration), it was mostly just a matter of program architecture and tying everything together.

Algorithm 1: Calculate Density

```
CalculateDensity
    inputs : set of windows W
    output : density ρ
    dataset: pseudomolecule C
    get genes from C
    get transposons from C
    initialize overlaps O
    foreach geneName g_j ∈ genes do
        foreach window w_j ∈ W do
            // see (1) (2) (3)
            o = Overlap(g_j, w_j, transposons)
            O[g_j, w_j] = o
    initialize densities ρ
    foreach geneName g_j ∈ genes do
        foreach TE identity t_j ∈ transposons do
            foreach window w_j ∈ W do
                // see (4)
                d = Density(g_j, t_j, w_j, O_j)
                ρ[g_j, w_j, t_j] = d
```

TE Density Software — Development

- For every gene, report the amount of TE-occupied base-pairs normalized by a given search space. I.e Gene X has 50% of the 5000 BP upstream area occupied by TEs, and then that number can further be broken down into the various TE types
- Software really isn't that complex (just a lot of iteration), it was mostly just a matter of program architecture and tying everything together.
- Insane 3D matrix to store all of the data (e.g $20 \times 15 \times 2$ for each gene of $108,000$ in strawberry)

Algorithm 1: Calculate Density

```
CalculateDensity
    inputs : set of windows W
    output : density ρ
    dataset: pseudomolecule C
    get genes from C
    get transposons from C
    initialize overlaps O
    foreach geneName g_j ∈ genes do
        foreach window w_j ∈ W do
            // see (1) (2) (3)
            o = Overlap(g_j, w_j, transposons)
            O[g_j, w_j] = o
    initialize densities ρ
    foreach geneName g_j ∈ genes do
        foreach TE identity t_j ∈ transposons do
            foreach window w_j ∈ W do
                // see (4)
                d = Density(g_j, t_j, w_j, O_j)
                ρ[g_j, w_j, t_j] = d
```

TE Density Software — Notable Improvements

- Speed — From a matter of weeks to 1-3 days

TE Density Software — Notable Improvements

- Speed — From a matter of weeks to 1-3 days
- Flexibility — Can be run on any genome

TE Density Software — Notable Improvements

- Speed — From a matter of weeks to 1-3 days
- Flexibility — Can be run on any genome
- Testing — Unit tests and integration tests

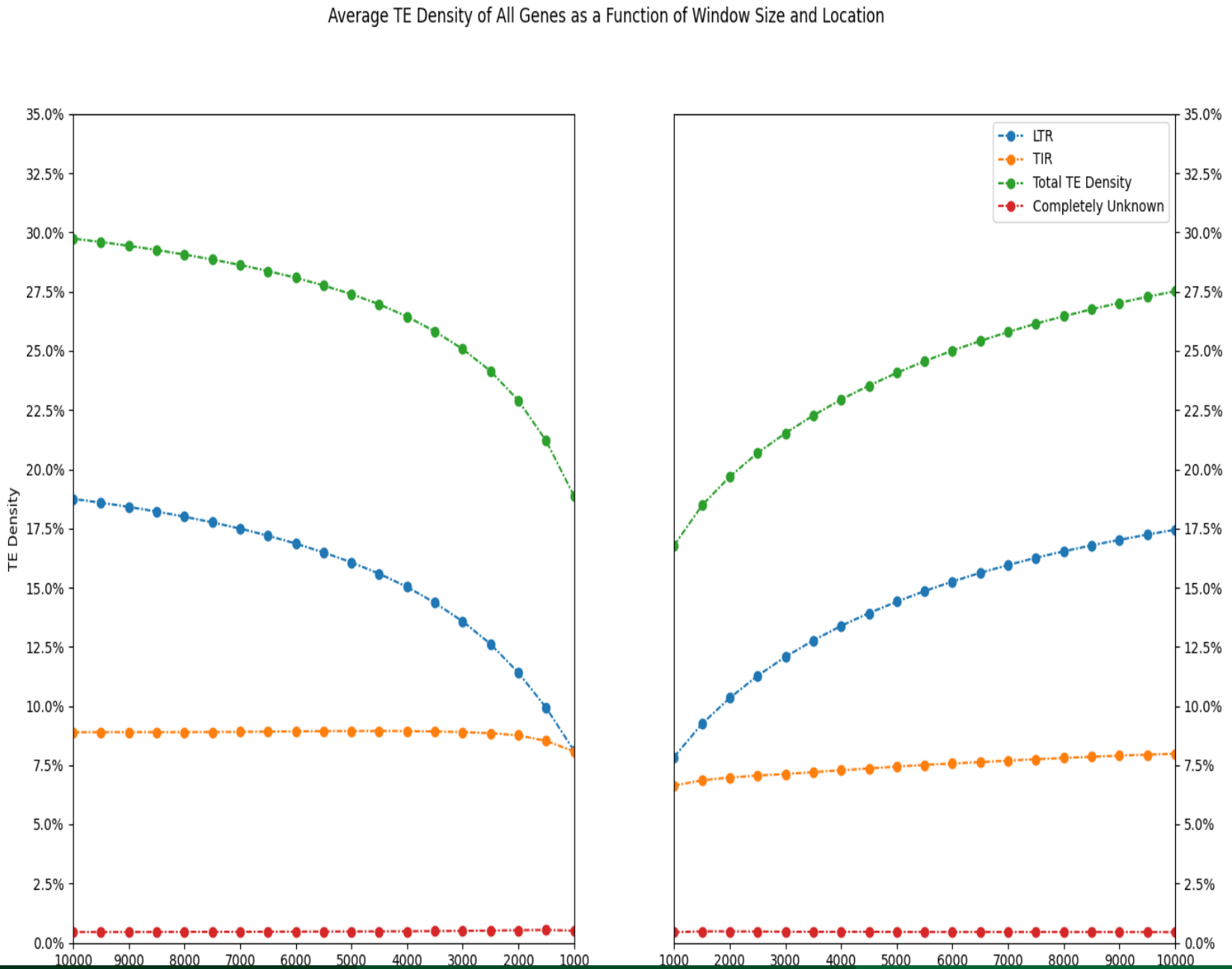
TE Density Software — Notable Improvements

- Speed — From a matter of weeks to 1-3 days
- Flexibility — Can be run on any genome
- Testing — Unit tests and integration tests
- Documentation — README and example figures

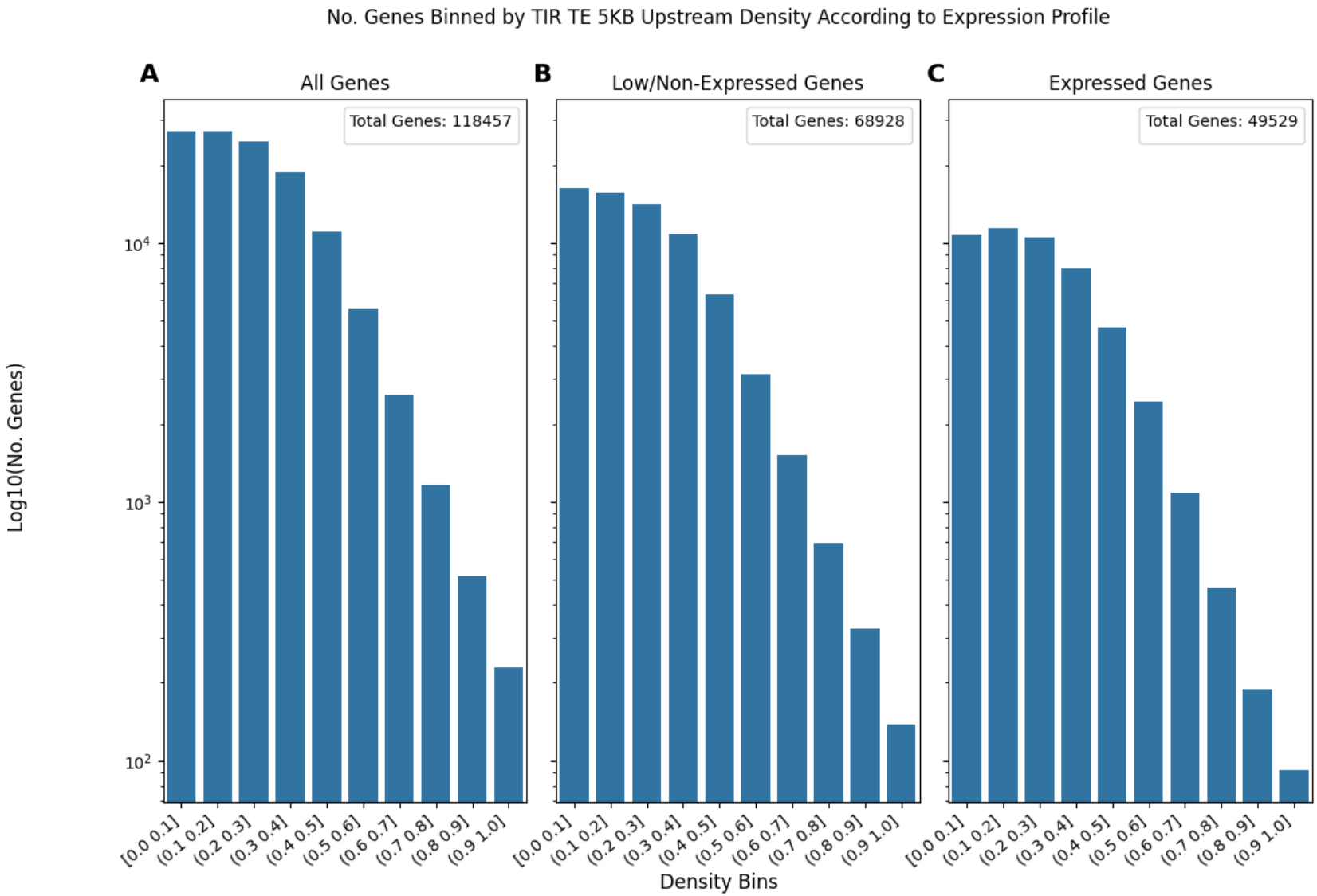
TE Density Software — Notable Improvements

- Speed — From a matter of weeks to 1-3 days
- Flexibility — Can be run on any genome
- Testing — Unit tests and integration tests
- Documentation — README and example figures
- Reproducibility — Python virtual environment

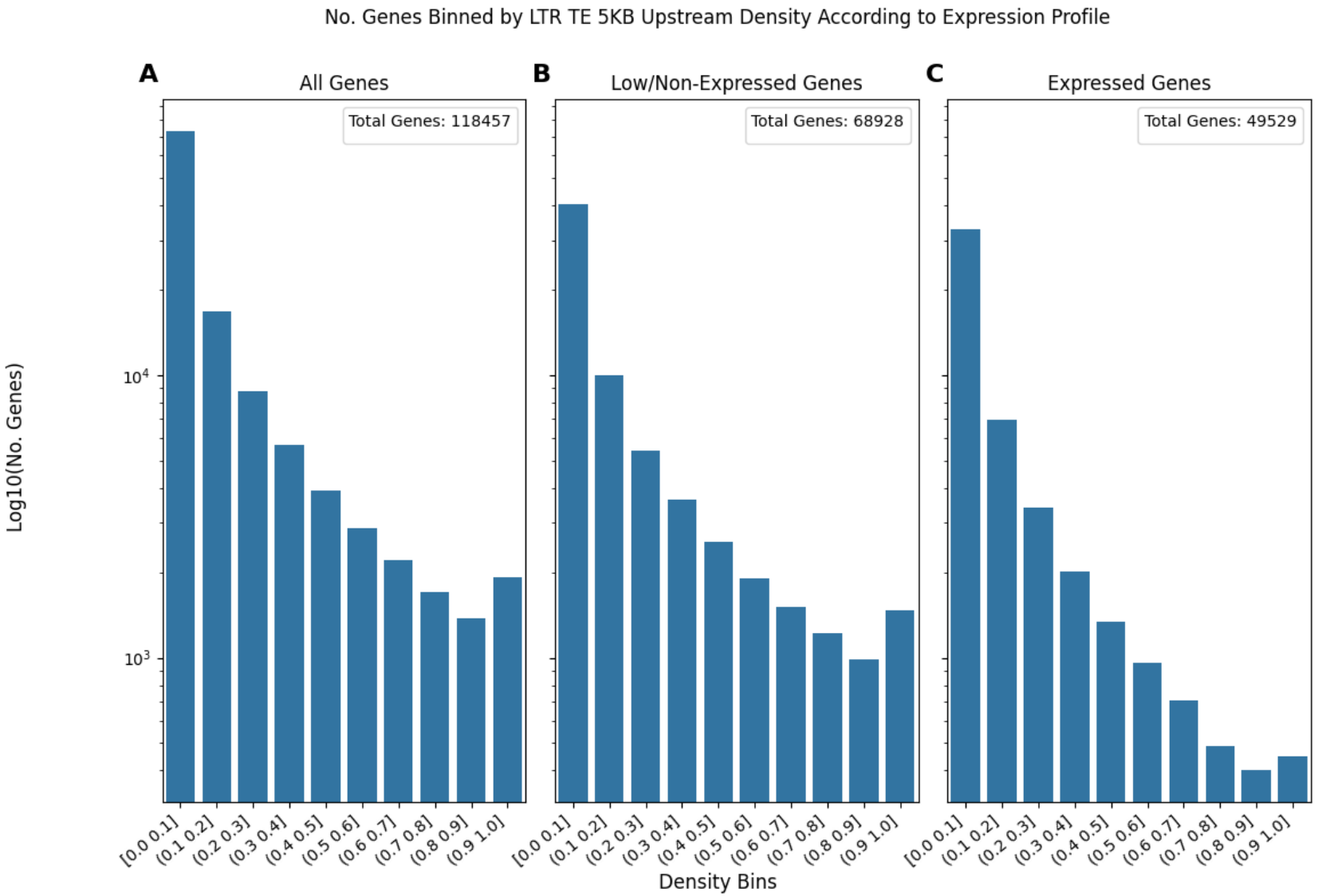
You can examine genome-wide trends of TE-gene relationships



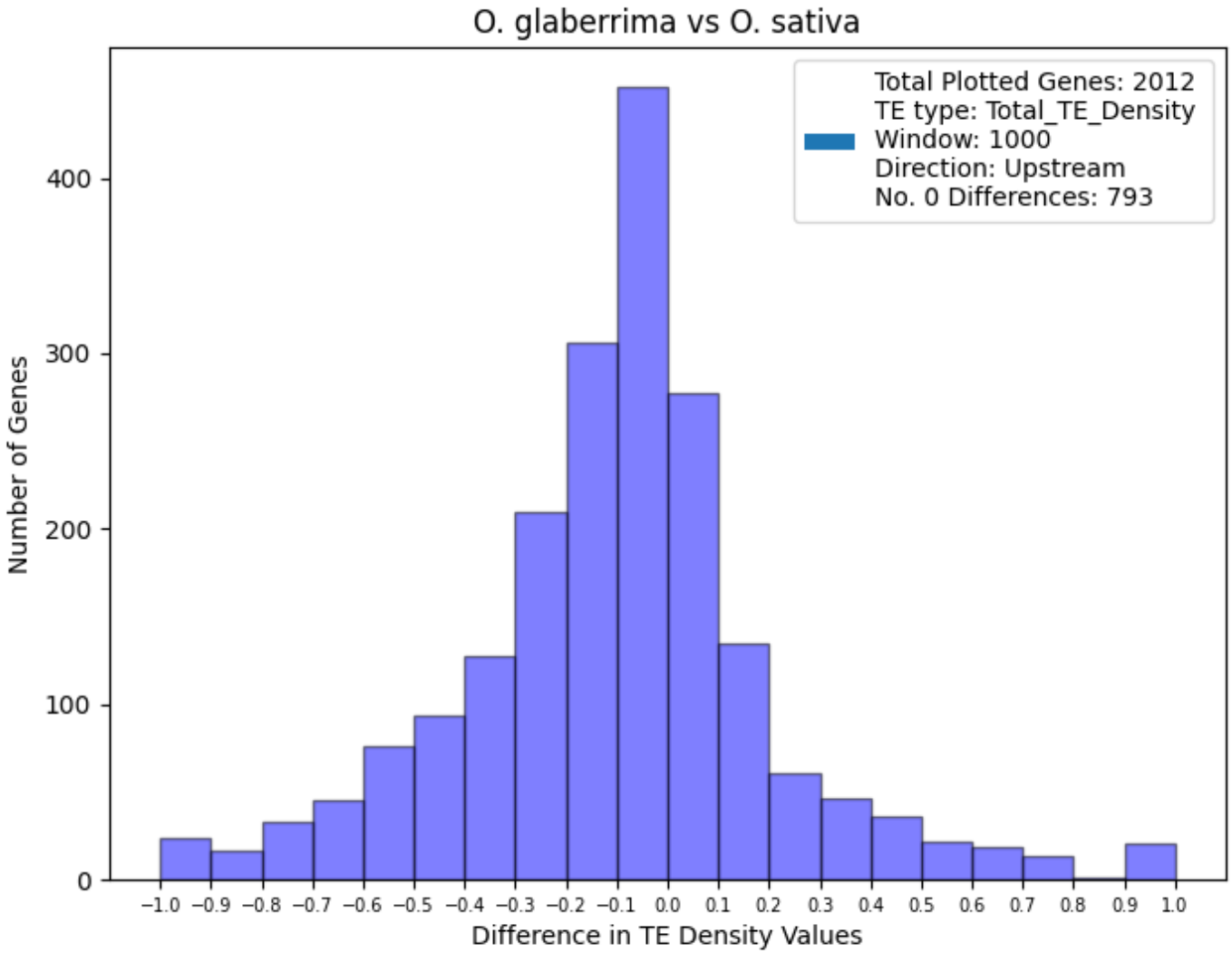
Re-examine the relationship with gene expression



Re-examine the relationship with gene expression



Identify TE differences between closely-related genes



What kinds of genes are associated with TEs?

More on this later!

PANTHER GO-Slim Biological Process	Fold Enrichment	FDR
Unclassified (UNCLASSIFIED)	1.11	0.00552
biological process (GO:0008150)	0.49	0.00276
cellular process (GO:0009987)	0.48	0.00687
organic substance metabolic process (GO:0071704)	0.46	0.0279
metabolic process (GO:0008152)	0.46	0.0175
primary metabolic process (GO:0044238)	0.44	0.0283
nitrogen compound metabolic process (GO:0006807)	0.42	0.0282
cellular metabolic process (GO:0044237)	0.42	0.0157

Strawberry TE Project — strawberry as a system

- Strawberries are octoploid (8 copies of each gene), and have a complex genome.

Strawberry TE Project — strawberry as a system

- Strawberries are octoploid (8 copies of each gene), and have a complex genome.
- Strawberries have a short history of domestication.

Strawberry TE Project — strawberry as a system

- Strawberries are octoploid (8 copies of each gene), and have a complex genome.
- Strawberries have a short history of domestication.
- Strawberries have a rich set of genomics resources, making it good system for studying polyploidy, domestication, and TE-gene relationships.

Strawberry is an extremely recent domesticate

- Cultivated strawberry (*Fragaria x. ananassa*) is a very recent domesticate, with the first cultivated strawberry appearing in the 18th century in the Royal Gardens of Versailles.

Sources: Edger et al (2019) - Nature Genetics & Darrow, G. M. (1966) - The Strawberry
Image: NC State Extension

Strawberry is an extremely recent domesticate

- Cultivated strawberry (*Fragaria x. ananassa*) is a very recent domesticate, with the first cultivated strawberry appearing in the 18th century in the Royal Gardens of Versailles.

- French Botanist Antoine Nicolas Duchesne is credited with its genesis; he created it by crossing the two New World species *Fragaria virginiana* and *Fragaria chiloensis*.

Sources: Edger et al (2019) - Nature Genetics & Darrow, G. M. (1966) - The Strawberry
Image: NC State Extension

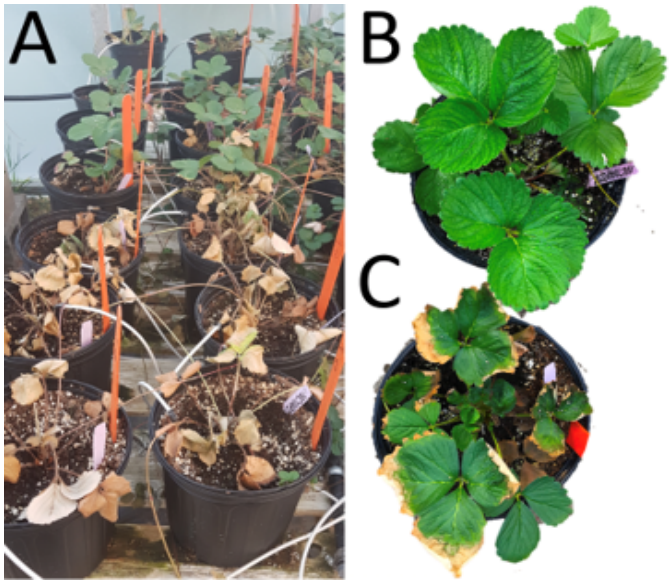
Strawberry TE Project

Given my findings from my REU, I am interested in understanding the role of TEs in strawberry domestication.

I do this by leveraging the cultivated strawberry genome as well as one of its wild progenitor species to compare and contrast TEs and genes.

A Comparable Wild Progenitor — *Fragaria chiloensis* "Del Norte"

- *Fragaria chiloensis* is a wild strawberry species that is a progenitor of the cultivated strawberry.
- It is also an octoploid species, and is quite hardy.
- It is found along the Pacific coast of North America.
- Our lab has sequenced and assembled the genome of this species, "Del Norte"



Strawberry TE Project

- **What were my questions?**

Strawberry TE Project

- **What were my questions?**
  - Do TEs shape phenotypes in strawberry?

Strawberry TE Project

- **What were my questions?**
  - Do TEs shape phenotypes in strawberry?
  - What kinds of genes are enriched with TEs? Are any of these genes associated with domestication phenotypes?

Strawberry TE Project

- **What were my questions?**
  - Do TEs shape phenotypes in strawberry?
  - What kinds of genes are enriched with TEs? Are any of these genes associated with domestication phenotypes?
  - What TE Density differences exist between wild and cultivated strawberries?

Strawberry TE Project

- **What were my questions?**
  - Do TEs shape phenotypes in strawberry?
  - What kinds of genes are enriched with TEs? Are any of these genes associated with domestication phenotypes?
  - What TE Density differences exist between wild and cultivated strawberries?
  - Can I generate a list of candidate TE-impacted genes for future study?

What datasets did I generate?

- TE pangenome with EDTA — What's that?
- Ortholog predictions
- TE Density calculations for each gene
- GO enrichments for TE-dense genes

Pangenomes — What and Why?

- Pangenomes are useful for understanding the genetic diversity of a species.

Pangenomes — What and Why?

- Pangenomes are useful for understanding the genetic diversity of a species.

- A pangenome is a collection of all the genes or TEs in a group of species, and it delineates the shared and unshared portions.

Pangenomes — What and Why?

- Pangenomes are useful for understanding the genetic diversity of a species.
- A pangenome is a collection of all the genes or TEs in a group of species, and it delineates the shared and unshared portions.
- TE pangenomes can be better than individual genome annotations because you have an enhanced ability to filter out false positives and negatives, particularly low-copy or fragmented TEs.

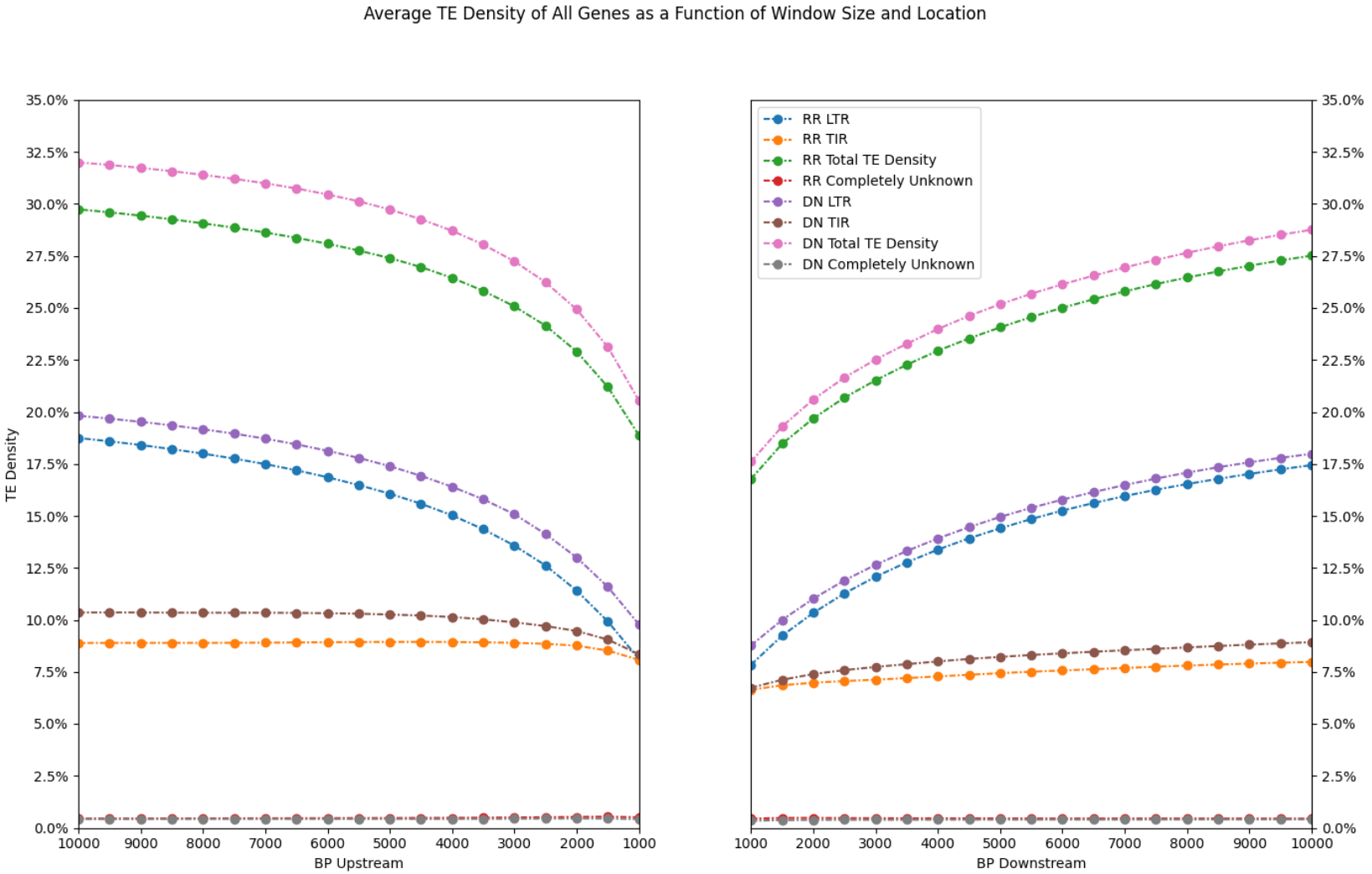
Pangenomes — What and Why?

- Pangenomes are useful for understanding the genetic diversity of a species.
- A pangenome is a collection of all the genes or TEs in a group of species, and it delineates the shared and unshared portions.
- TE pangenomes can be better than individual genome annotations because you have an enhanced ability to filter out false positives and negatives, particularly low-copy or fragmented TEs.
- I wanted a consistent and comparable TE annotation for both species, where I could compare and contrast differing TE families (which are defined by sequence similarity).

TE pangenome reveals subtle differences between wild and cultivated strawberry

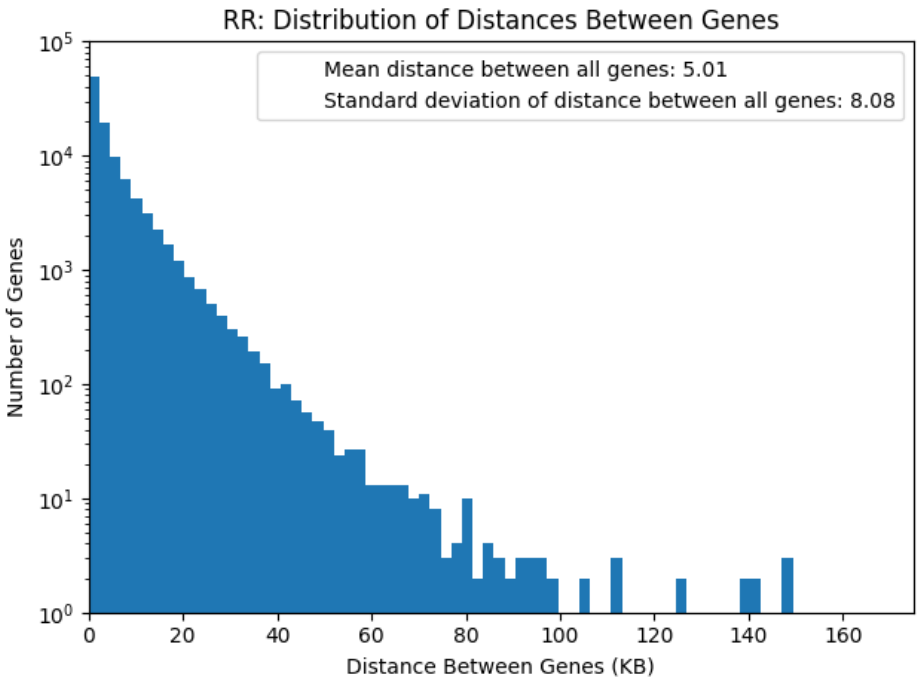
Class	Subtype	H4	DN	RR
LINE	L1	0.57%	0.46%	0.47%
LTR	Copia	3.46%	3.67%	3.51%
	Gypsy	5.05%	7.75%	7.90%
	unknown	7.62%	13.72%	12.57%
SINE	tRNA	0.00%	0.00%	0.00%
	unknown	0.37%	0.48%	0.56%
TIR	CACTA (En/Spm)	3.71%	4.88%	3.86%
	Mutator	2.43%	2.35%	2.16%
	PIF Harbinger	0.97%	0.91%	0.91%
	Tc1 Mariner	0.04%	0.02%	0.03%
	hAT	1.85%	2.16%	1.93%
Low Complexity	None	0.01%	0.07%	0.02%
nonLTR	pararetrovirus	0.13%	0.17%	0.16%
nonTIR	helitron	1.87%	1.17%	1.24%
rDNA	45S	0.01%	0.01%	0.01%
Repeat Fragment	None	0.54%	0.45%	0.56%
Total		**28.62%**	**38.28%**	**35.88%**

Cultivated strawberry is less TE-dense near genes

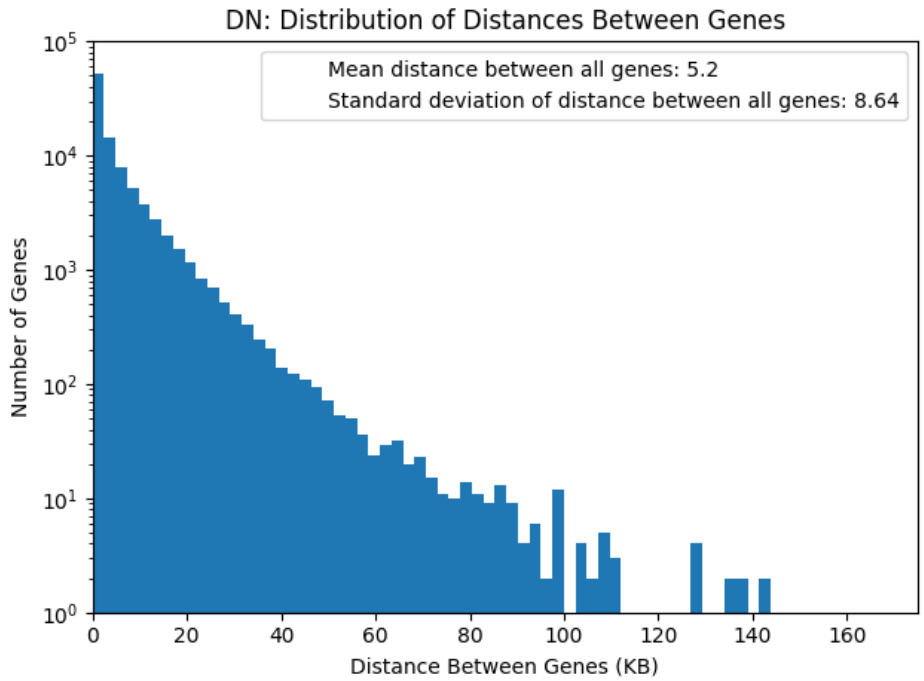


Cultivated strawberry is more gene-dense

Cultivated



Wild



What are the functional enrichments of TE-dense genes? How much do they differ?

- **What were my questions?**
  - Do TEs shape phenotypes in strawberry?
  - What TE Density differences exist between wild and cultivated strawberries?
  - Can I generate a list of candidate TE-impacted genes for future study?
  - What kinds of genes are enriched with TEs? Are any of these genes associated with domestication phenotypes?

What are the functional enrichments of TE-dense genes? How much do they differ?

GO Term	P-Value
Unique to Wild Strawberry	
Pectic galactan metabolic process	0.04842
Starch catabolic process	0.00785
Peptidoglycan biosynthetic process	0.04842
Tricarboxylic acid cycle	0.03893
Leaf development	0.00014
Embryo development ending in seed dormancy	0.00697
Regulation of developmental process	0.01241
Maltose biosynthetic process	0.04842
Malate metabolic process	0.00971
Circadian regulation of gene expression	0.01794
Histone H3-K9 demethylation	0.00342
Plant-type secondary cell wall biogenesis	0.00051
Cellular water homeostasis	0.006800
Pigment metabolic process	0.02453
Plant ovule development	0.02538
Negative regulation of flower development	0.04217

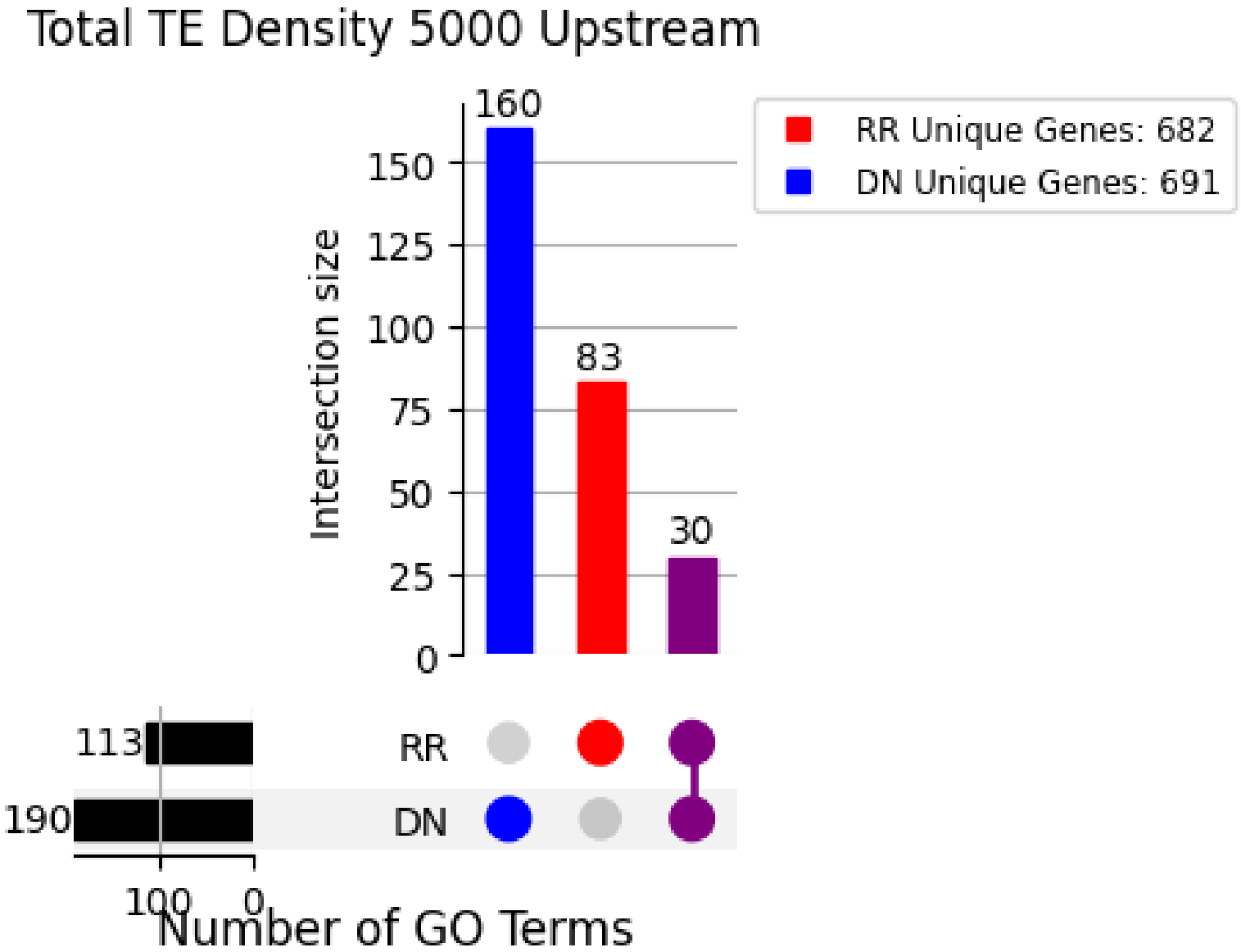
What are the functional enrichments of TE-dense genes? How much do they differ?

GO Term	P-Value
Unique to Domesticated Strawberry	
Fructose 6-phosphate metabolic process	0.00776
Brassinosteroid homeostasis	0.03174
Response to flooding	0.02728
Regulation of jasmonic acid biosynthetic process	0.03944
Abscisic acid homeostasis	0.00882
Reponse to microbial phytotoxin	0.03944
Flower development	0.01750
Seed development	0.01059
Shoot apical meristem development	0.02723
Glycine betain biosynthetic process from choline	0.00882
UDP-glucose metabolic process	0.02725

What are the functional enrichments of TE-dense genes? How much do they differ?

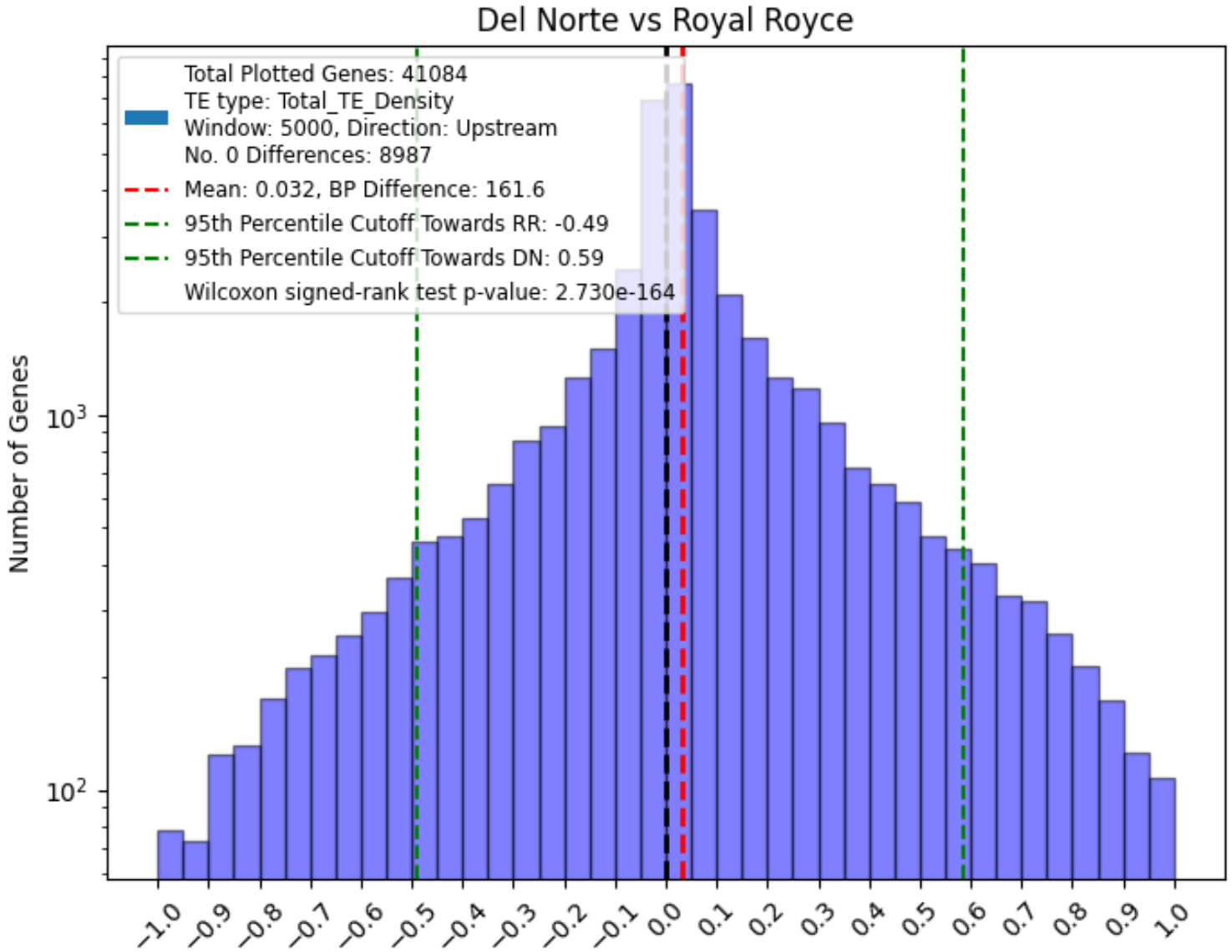
GO Term	P-Value	
Shared	**Wild vs. Domesticated**	
Cellular defense response	0.00234	0.00305
Response to osmotic stress	0.02834	0.02382
Response to heat	0.01600	0.01904
Gibberellin biosynthetic process	0.00623	0.04642
Regulation of gene expression	0.00440	0.00440
Vegetative to Reproductive Phase Transition of Meristem	0.00883	0.00592

What are the functional enrichments of TE-dense genes? How much do they differ?



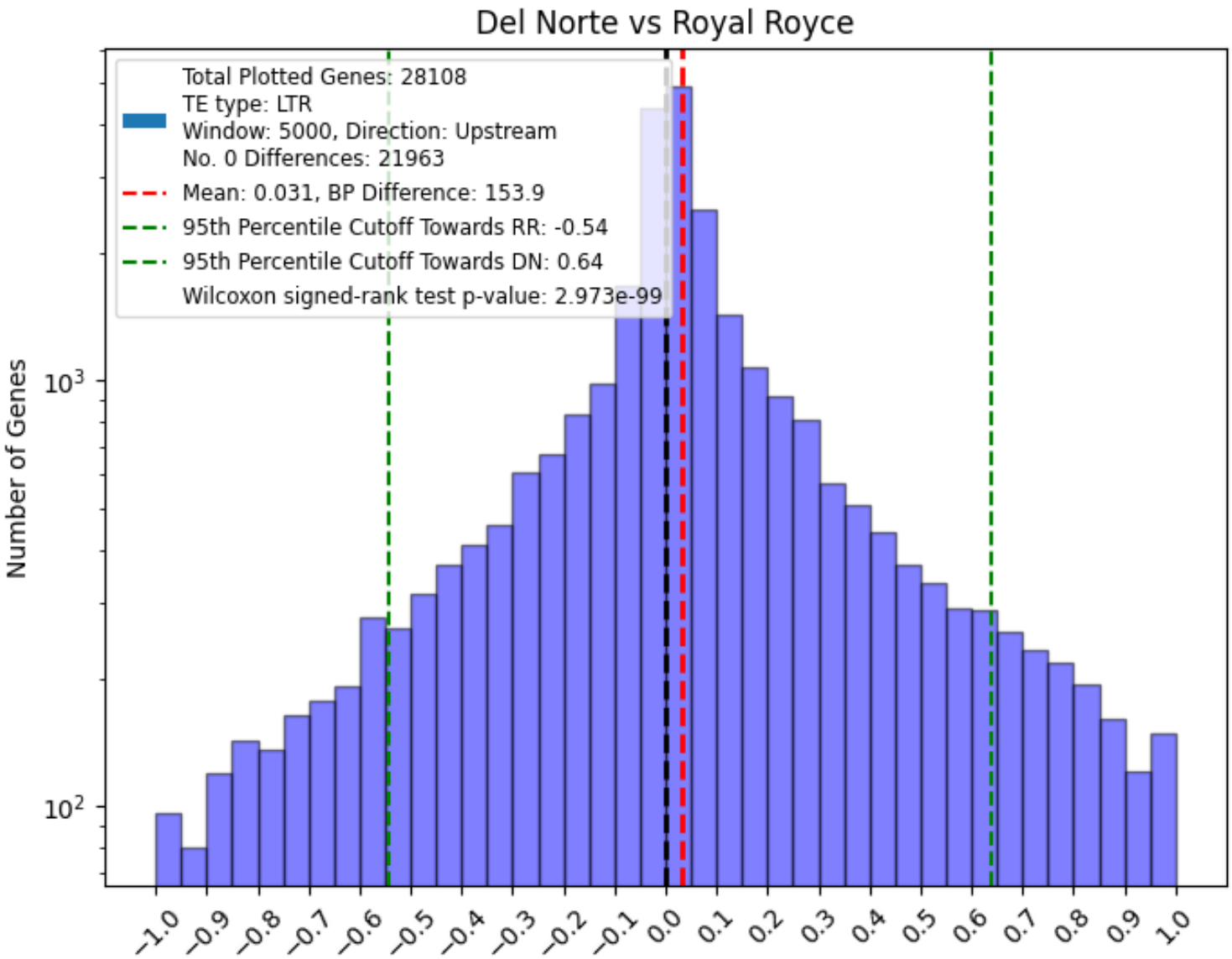
Syntelogs show major differences in TE density

All TEs



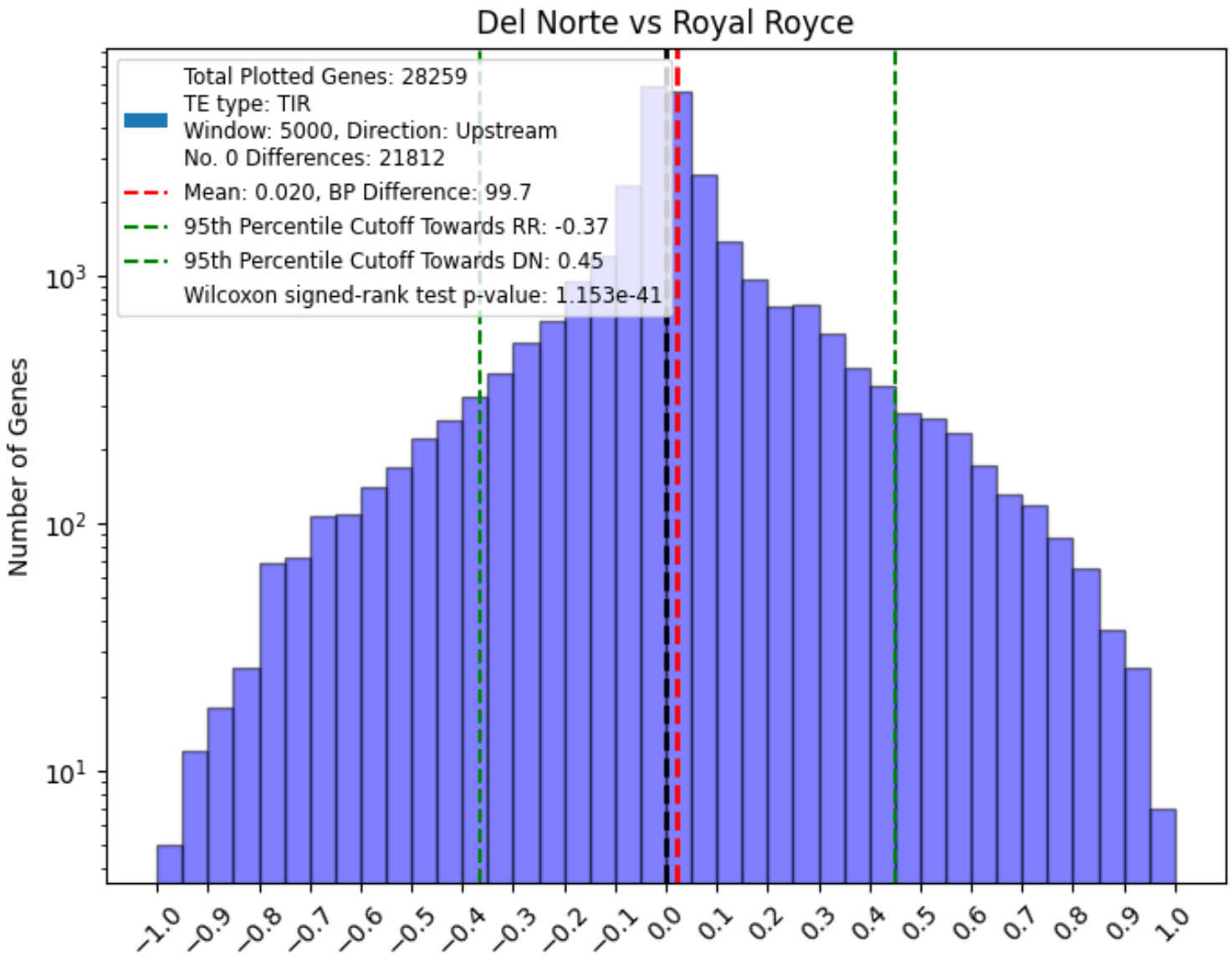
Syntelogs show major differences in TE density

LTR TEs



Syntelogs show major differences in TE density

TIR TEs



Syntelogs show major differences in TE density

Cultivated strawberry has a less TEs around positionally conserved genes.

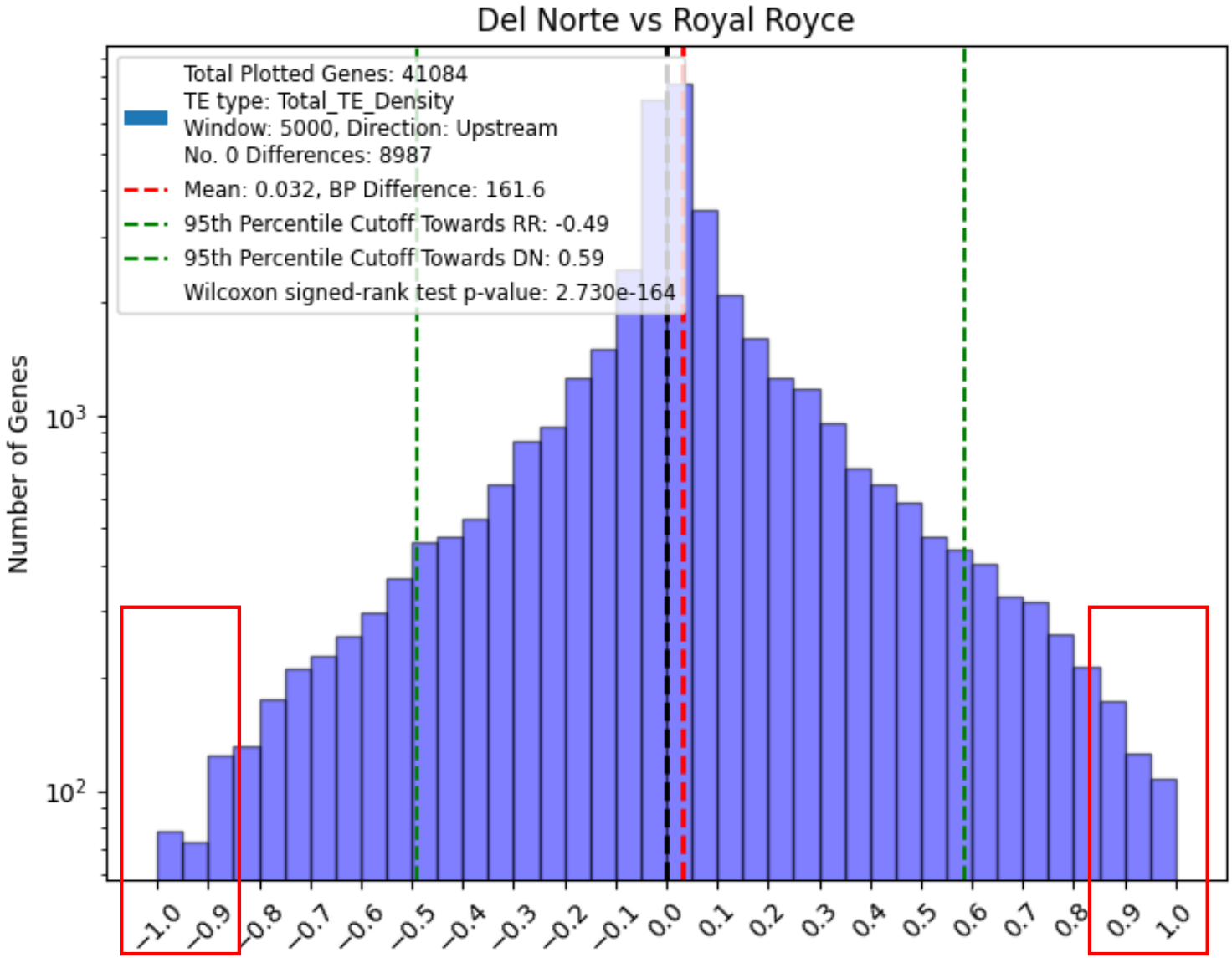
- For every TE type
- For every measurement window
- For every direction – upstream/downstream

Syntelogs show major differences in TE density

TE Category	Window	BP Bias	P-Value
Total TE	2500	113.7	1.603e-228
	5000	161.6	2.730e-164
	7500	202.3	4.871e-137
LTR	2500	125.6	5.776e-141
	5000	153.9	2.973e-99
	7500	171.9	7.172e-79
TIR	2500	73.2	4.140e-51
	5000	99.7	1.153e-41
	7500	126.7	5.753e-49
TIR/CACTA	2500	148.7	3.844e-33
	5000	218.0	1.813e-41
	7500	250.0	1.368e-48

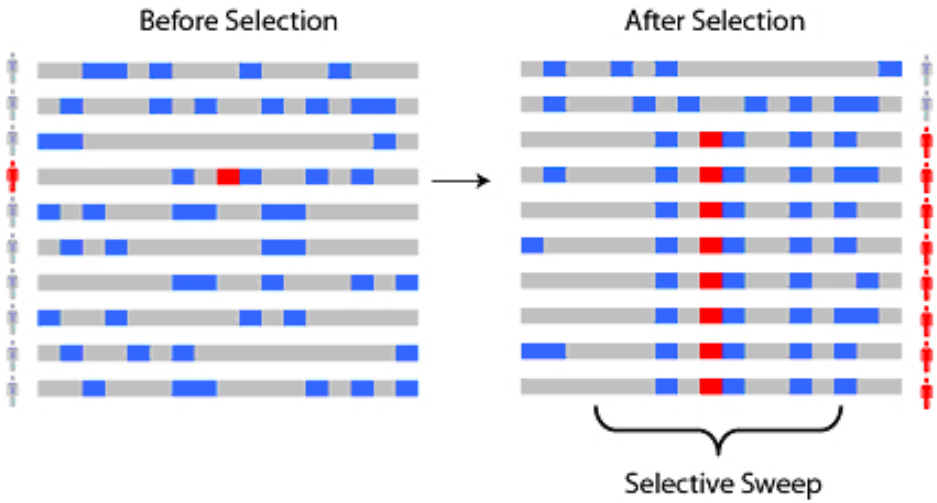
Table: Summary table of TE density differences between Del Norte and Royal Royce syntelogs. The BP Bias column reflects the average base-pair bias towards Del Norte, and the P-Value column reflects the significance of the Wilcoxon signed-rank test.

What kinds of genes differ in their TE presence between wild and cultivated strawberry?



Unique TE insertions in selective sweeps in cultivated strawberry

- The process of domestication entails the selection and preferential propagation of specific phenotypes, which leaves signatures of selection in the genome[1].

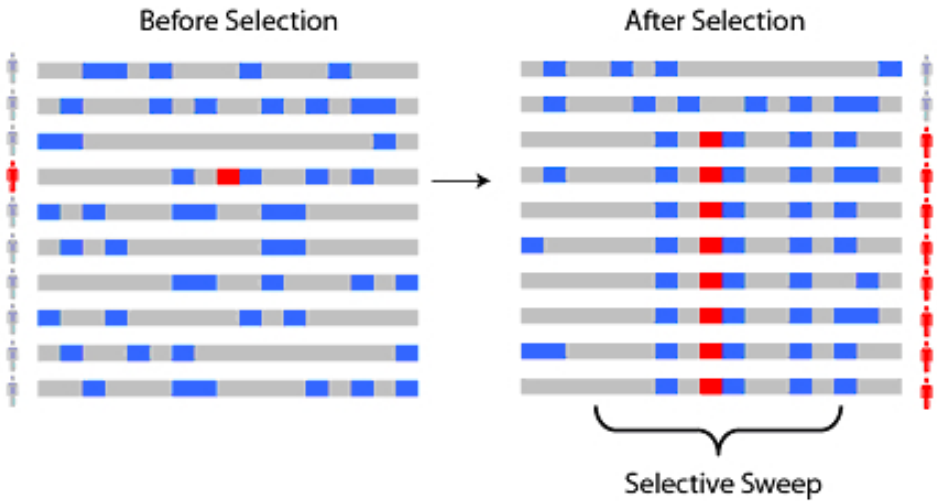


[1] Hardigan et al. (2021) - Molecular Biology and Evolution

Unique TE insertions in selective sweeps in cultivated strawberry

- The process of domestication entails the selection and preferential propagation of specific phenotypes, which leaves signatures of selection in the genome[1].

- Selective sweeps are created when selection decreases genetic variation among neutral loci that are linked to the selected locus.



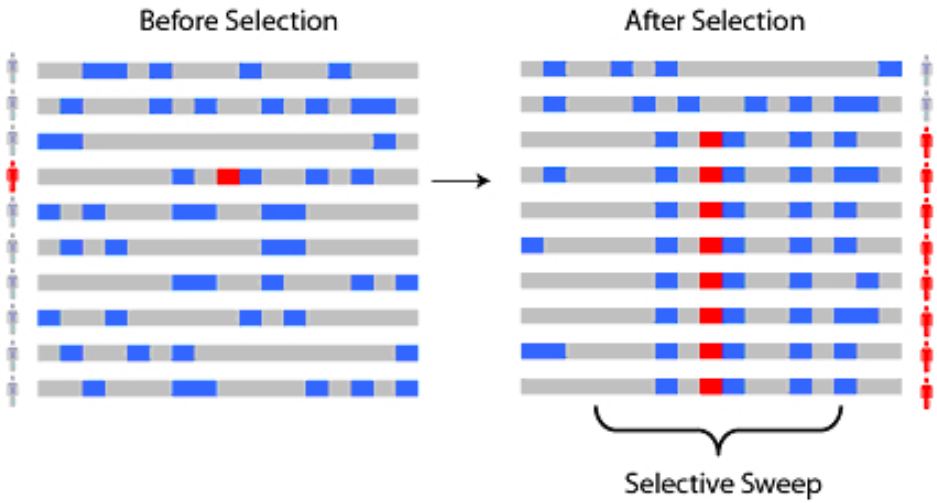
[1] Hardigan et al. (2021) - Molecular Biology and Evolution

Unique TE insertions in selective sweeps in cultivated strawberry

- The process of domestication entails the selection and preferential propagation of specific phenotypes, which leaves signatures of selection in the genome[1].

- Selective sweeps are created when selection decreases genetic variation among neutral loci that are linked to the selected locus.

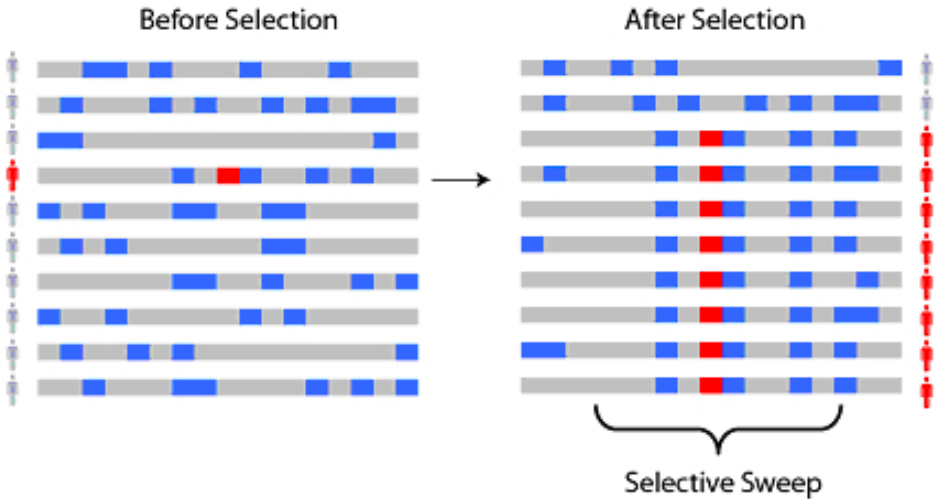
- Selective sweeps are regions of the genome where a beneficial allele has rapidly increased in frequency, and the region has reduced genetic diversity.



[1] Hardigan et al. (2021) - Molecular Biology and Evolution

Unique TE insertions in selective sweeps in cultivated strawberry

- I used data from previous publications[1] to intersect my differing TE-dense genes with regions that we know are associated with early and modern domestication.

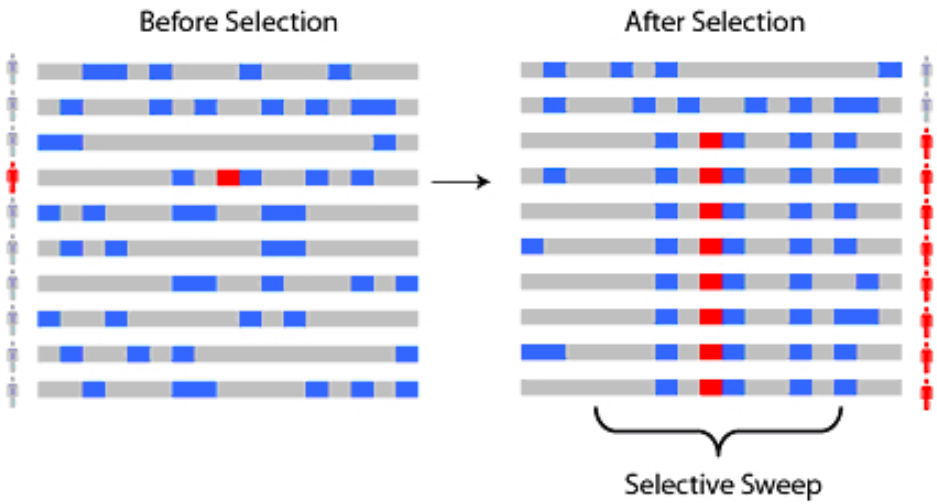


[1] Fan and Whitaker (2024) - Plant Cell

Unique TE insertions in selective sweeps in cultivated strawberry

- I used data from previous publications[1] to intersect my differing TE-dense genes with regions that we know are associated with early and modern domestication.

- I identified 60 genes residing within selective sweep regions that had significant functional enrichments and a TE density difference greater than or equal to 0.75.



What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

- **Why do I think TEs are involved?**

What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

- **Why do I think TEs are involved?**
  - They are all in selective sweep regions, early and/or modern domestication zones

What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

- **Why do I think TEs are involved?**
  - They are all in selective sweep regions, early and/or modern domestication zones
  - There is no analogous TE in the wild strawberry genome, TE is unique to cultivated strawberry

What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

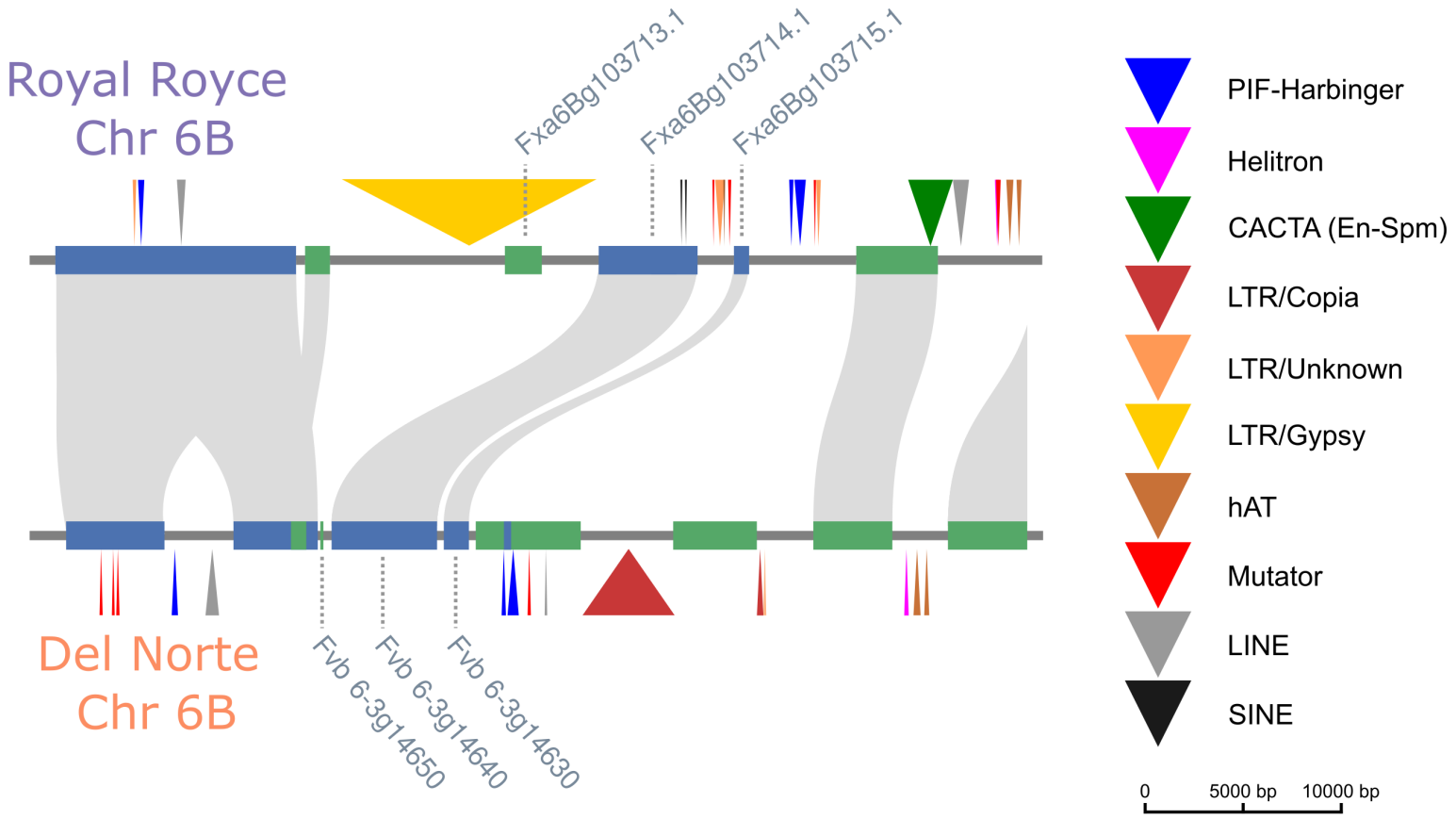
- **Why do I think TEs are involved?**
  - They are all in selective sweep regions, early and/or modern domestication zones
  - There is no analogous TE in the wild strawberry genome, TE is unique to cultivated strawberry
  - The TE is immediately upstream of the gene, and is not decayed at all

What notable genes did I find?

- Fructose 6 phosphate gene — sugar metabolism
- A SWEET gene — sugar transport
- COP1 Supressor 2 — light signaling and development

- **Why do I think TEs are involved?**
  - They are all in selective sweep regions, early and/or modern domestication zones
  - There is no analogous TE in the wild strawberry genome, TE is unique to cultivated strawberry
  - The TE is immediately upstream of the gene, and is not decayed at all
  - The TE is in a family with relatively low copy number, suggesting it is rather new

Fructose


Royal Royce
Chr 6B
Fxa6Bg103713.1
Fxa6Bg103714.1
Fxa6Bg103715.1
Del Norte
Chr 6B
Fvb 6-3g14650
Fvb 6-3g14640
Fvb 6-3g14630
PIF-Harbinger
Helitron
CACTA (En-Spm)
LTR/Copia
LTR/Unknown
LTR/Gypsy
hAT
Mutator
LINE
SINE
0
5000 bp
10000 bp

Summary

- I applied TE Density software to strawberries, and found that cultivated strawberry has overall less TEs, and that it has less TEs near genes for every TE type and measurement parameter.

Summary

- I applied TE Density software to strawberries, and found that cultivated strawberry has overall less TEs, and that it has less TEs near genes for every TE type and measurement parameter.

- Functional enrichment analyses point to distinct patterns of TE-associated genes between the two genomes, particularly in pathways relevant to strawberry breeding.

Summary

- I applied TE Density software to strawberries, and found that cultivated strawberry has overall less TEs, and that it has less TEs near genes for every TE type and measurement parameter.
- Functional enrichment analyses point to distinct patterns of TE-associated genes between the two genomes, particularly in pathways relevant to strawberry breeding.
- I identified a number of fruit, defense, transcription factor, and development genes enriched for novel TE presence in domesticated strawberry, demonstrating the potential for TEs to influence a variety of important domestication related traits.

Summary

- I applied TE Density software to strawberries, and found that cultivated strawberry has overall less TEs, and that it has less TEs near genes for every TE type and measurement parameter.
- Functional enrichment analyses point to distinct patterns of TE-associated genes between the two genomes, particularly in pathways relevant to strawberry breeding.
- I identified a number of fruit, defense, transcription factor, and development genes enriched for novel TE presence in domesticated strawberry, demonstrating the potential for TEs to influence a variety of important domestication related traits.
- The TE Density software offers a powerful method to identify genes that are potentially impacted by TEs

Summary

- I applied TE Density software to strawberries, and found that cultivated strawberry has overall less TEs, and that it has less TEs near genes for every TE type and measurement parameter.
- Functional enrichment analyses point to distinct patterns of TE-associated genes between the two genomes, particularly in pathways relevant to strawberry breeding.
- I identified a number of fruit, defense, transcription factor, and development genes enriched for novel TE presence in domesticated strawberry, demonstrating the potential for TEs to influence a variety of important domestication related traits.
- The TE Density software offers a powerful method to identify genes that are potentially impacted by TEs
- Most of the time, when someone identifies a TE-influenced gene, it is because they first identified an unusual phenotype, *and then* they identified the gene, *and then* saw a TE was sitting right next to it. My methodology can be used to start with the genotype and work forwards.

Future Directions

- We know that TEs can influence gene expression, how many of the genes are being up vs. downregulated by a TE?

Future Directions

- We know that TEs can influence gene expression, how many of the genes are being up vs. downregulated by a TE?

- If a TE is altering the expression of something like a hormone gene, that can have far reaching effects. How can we analyze the downstream effects?

Future Directions

- We know that TEs can influence gene expression, how many of the genes are being up vs. downregulated by a TE?
- If a TE is altering the expression of something like a hormone gene, that can have far reaching effects. How can we analyze the downstream effects?
- What are the tissue-specific or temporal expression contexts of my candidate genes?

Future Directions

- We know that TEs can influence gene expression, how many of the genes are being up vs. downregulated by a TE?
- If a TE is altering the expression of something like a hormone gene, that can have far reaching effects. How can we analyze the downstream effects?
- What are the tissue-specific or temporal expression contexts of my candidate genes?
- Why do I have a fructose gene *inside* a TE (an LTR element no less!)? Evidence suggests it is a real gene, TE seems real too. Where did they both come from?

Future Directions

- We know that TEs can influence gene expression, how many of the genes are being up vs. downregulated by a TE?
- If a TE is altering the expression of something like a hormone gene, that can have far reaching effects. How can we analyze the downstream effects?
- What are the tissue-specific or temporal expression contexts of my candidate genes?
- Why do I have a fructose gene *inside* a TE (an LTR element no less!)? Evidence suggests it is a real gene, TE seems real too. Where did they both come from?
- What are the subgenome-specific patterns of TE insertions, what does the expression and density of the other homeologs look like?

Collaborations

Collaborations

I am teaming up with the Batman and the Antwoman

Collaborations

David Ray — The Batman

BAT 1K

- Dr. David Ray of Texas Tech
- Bats have unusually long lifespans, given their body size (Longest living wild bat was 41 years old).
- Comparative analysis of bats and rodents' genomes suggests a relation between non-LTR retrotransposons, cancer incidence, and ageing[1]

[1] Ricci et al. (2023) - Scientific Reports

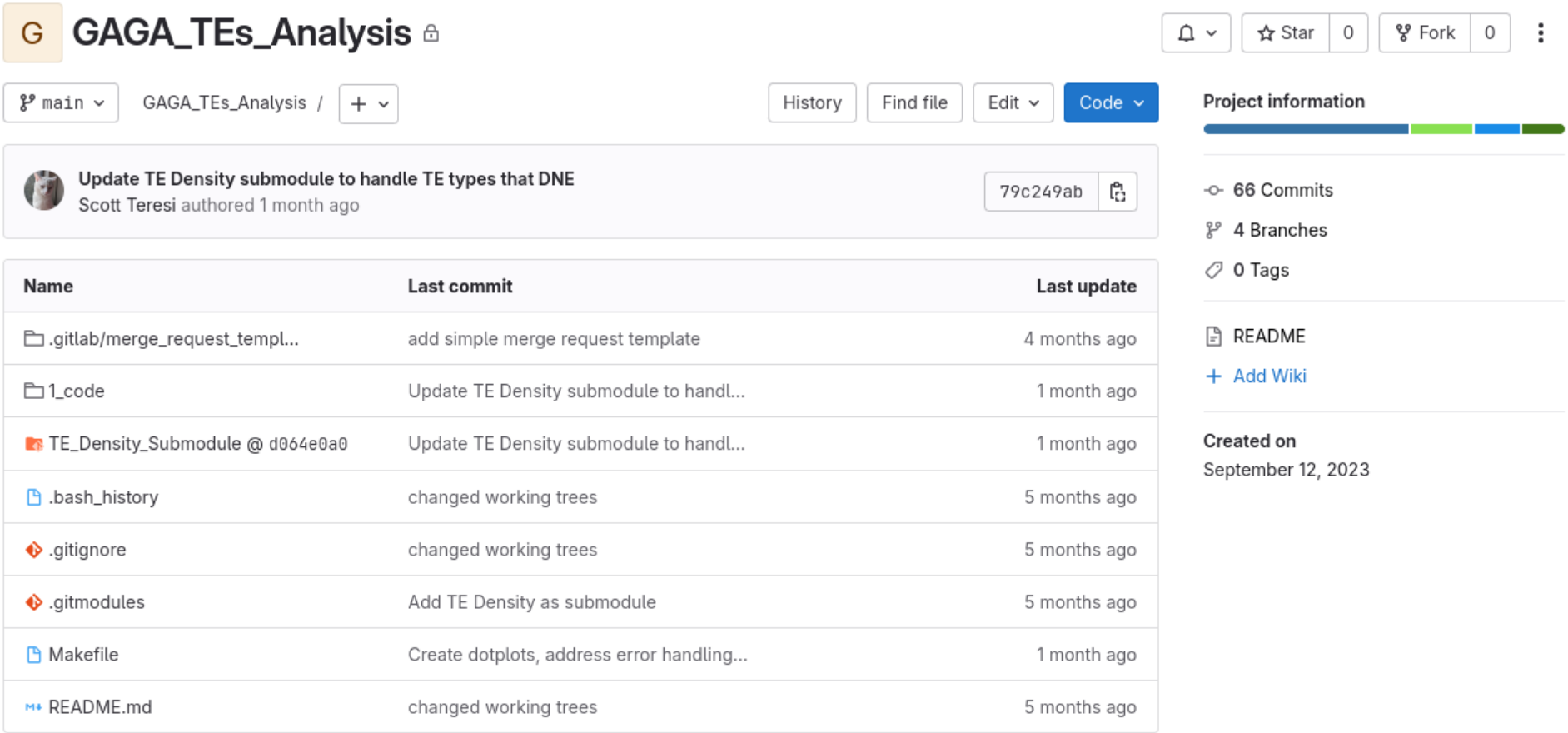
Collaborations

Janina Rinke — The Antwoman

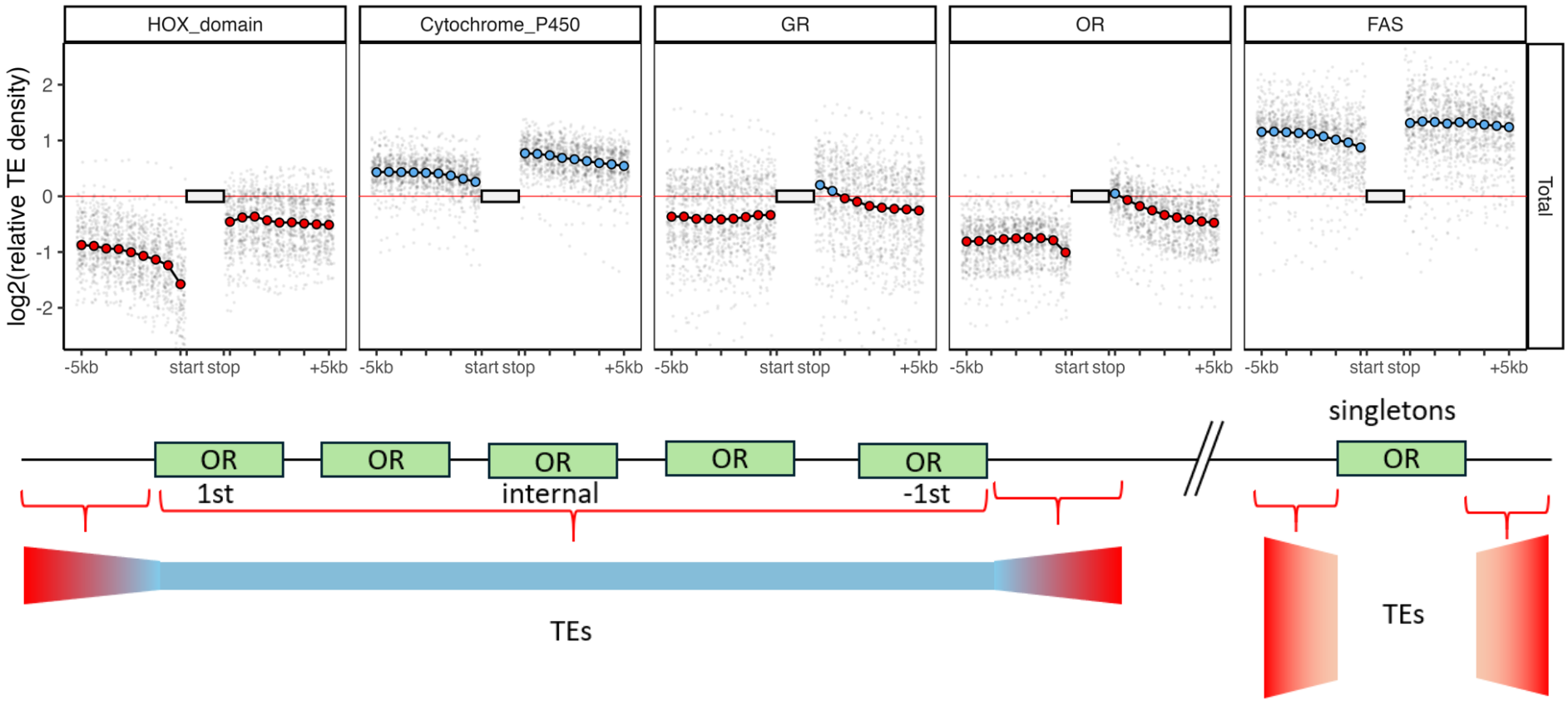
- The Global Ant Genomics Alliance!
- Janina Rinke and Dr. Lukas Schrader of the University of Münster
- Ants are social insects and have evolved complex mechanisms for communication and chemical sensing

Collaborations

Janina Rinke — The Antwoman



Preliminary results show depletion/enrichment of TEs near certain gene families, as well as an association between TE abundance and gene family expansions/contractions.

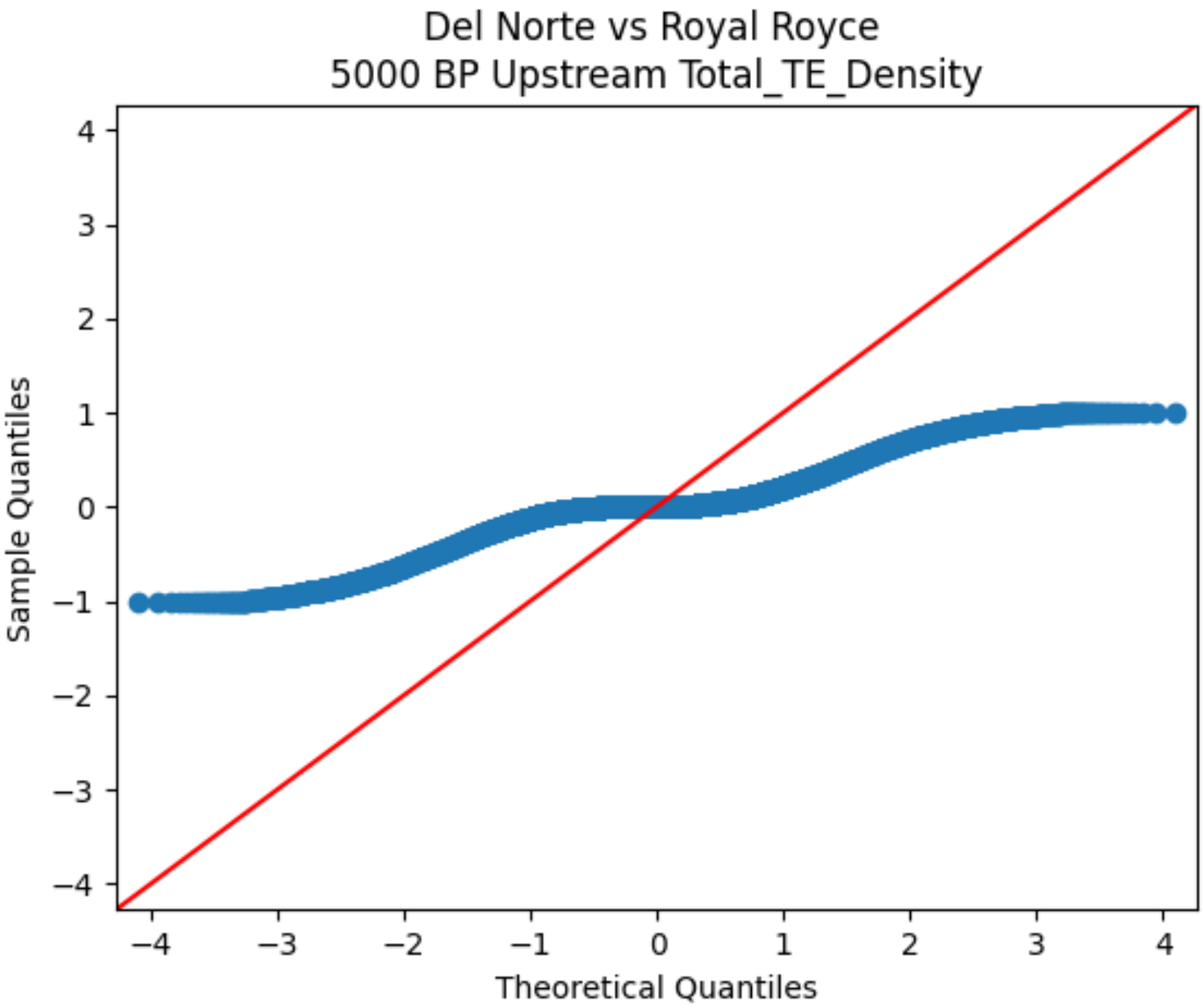


Acknowledgements

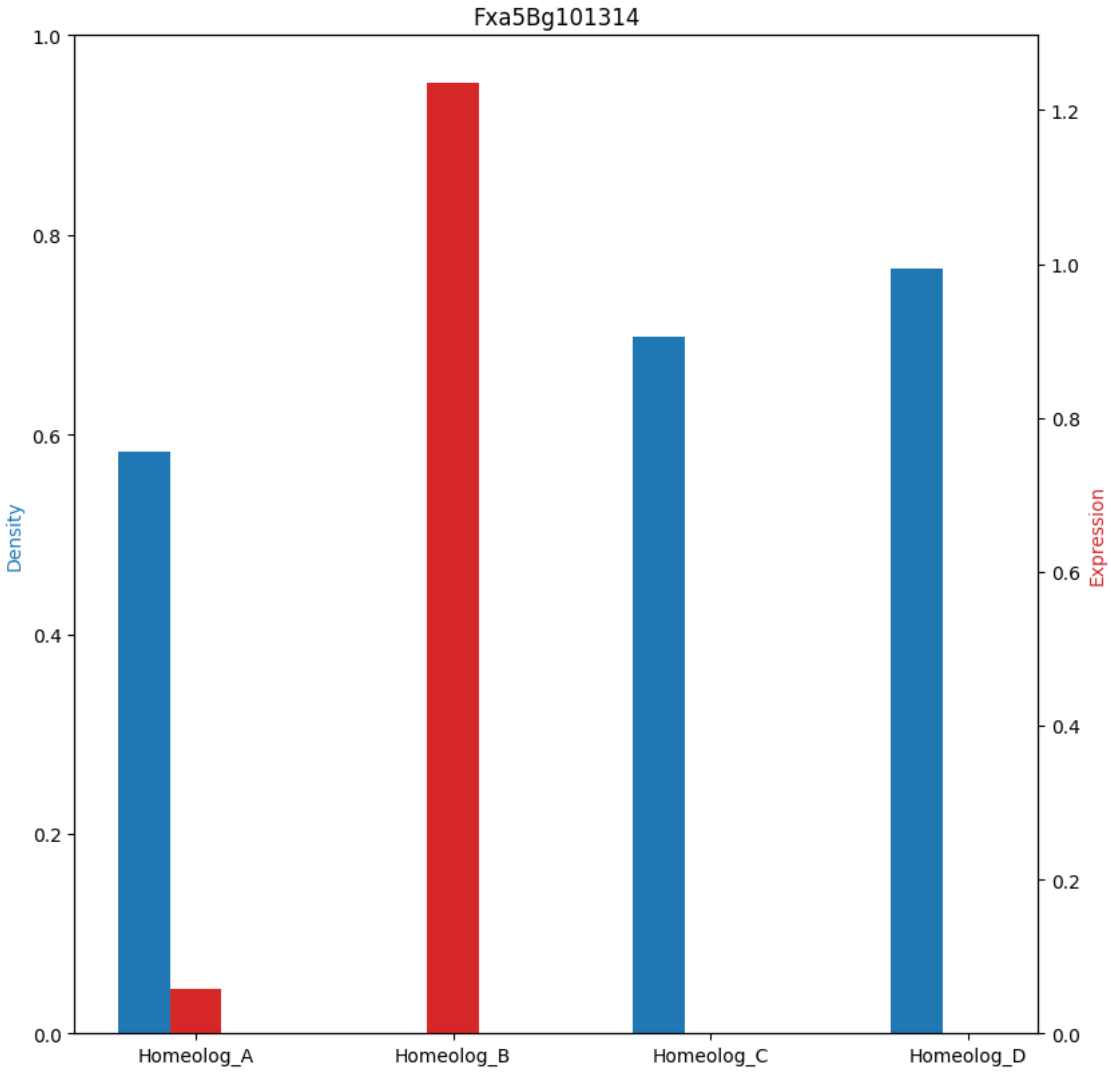
- Jay & Pam Teresi, Michael, Steven, Jane
- Erin Kramer
- Pat Edger & the Edger Lab
- Ning Jiang, Jiming Jiang, Jianrong Wang, Shin-Han Shiu
- Scott Stelpflug, Sarah Jensen, Harly Durbin-Rowan
- Shujun Ou
- Ryan Williams
- Jyothi Kumar
- Meghan Hill
- Alder Fulton
- Plant Genomics REU
- Ross Hatlen, Mallory St. Clair, Jordan Brock, Michael Gasdick
- James Bowers, Brandon Hinton, Drew Reibel
- Dan Wyrembelski
- Nicholas Panchy
- Joshua Puzey

Questions?

Extras


Del Norte vs Royal Royce
5000 BP Upstream Total_TE_Density
Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

Extras


Fxa5Bg101314
Density
Expression
1.0
0.8
0.6
0.4
0.2
0.0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Homeolog_A
Homeolog_B
Homeolog_C
Homeolog_D