Gillenia Genome V2

Cecilia Deng, Ting-Hsuan Chen, David Chagne

Gillenia Genome Work

- INRAE: assembled good quality contigs
- PFR: Reference-guided scaffolding
- TE detection and repeat masking

Chr-level Scaffolding: Hap1

- Hap1 is done and the 9 chrs can be aligned to our previous NCBI Gillenia assembly quite well.

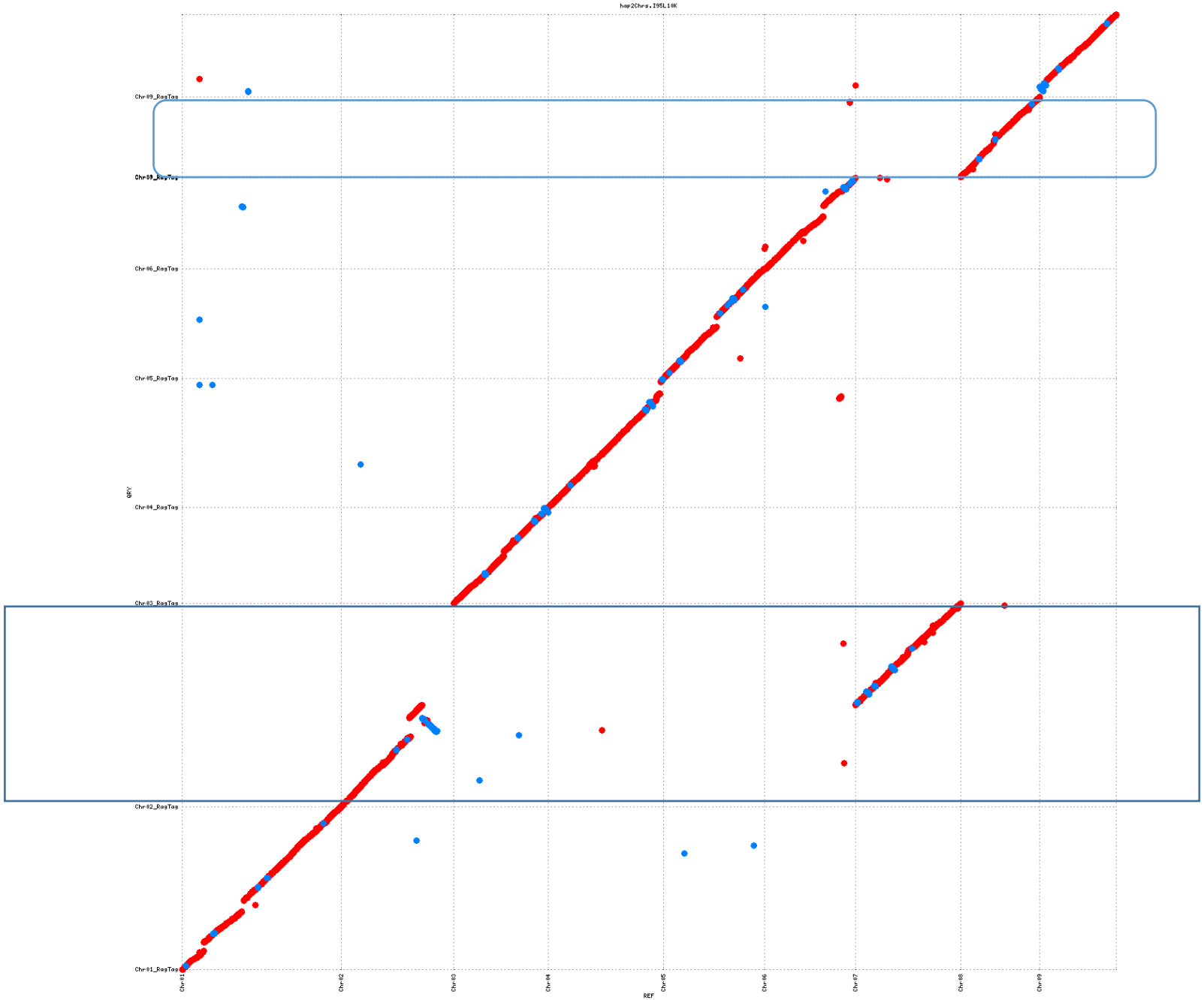
- There are some spaces to further improve the assembly if more time/data are available:

  - H1Chr 01, 04 and 09 only have telomere peak at one end
  - H1Chr06 and 07 have a wider telomere peak at 3’ and 5’, respectively.

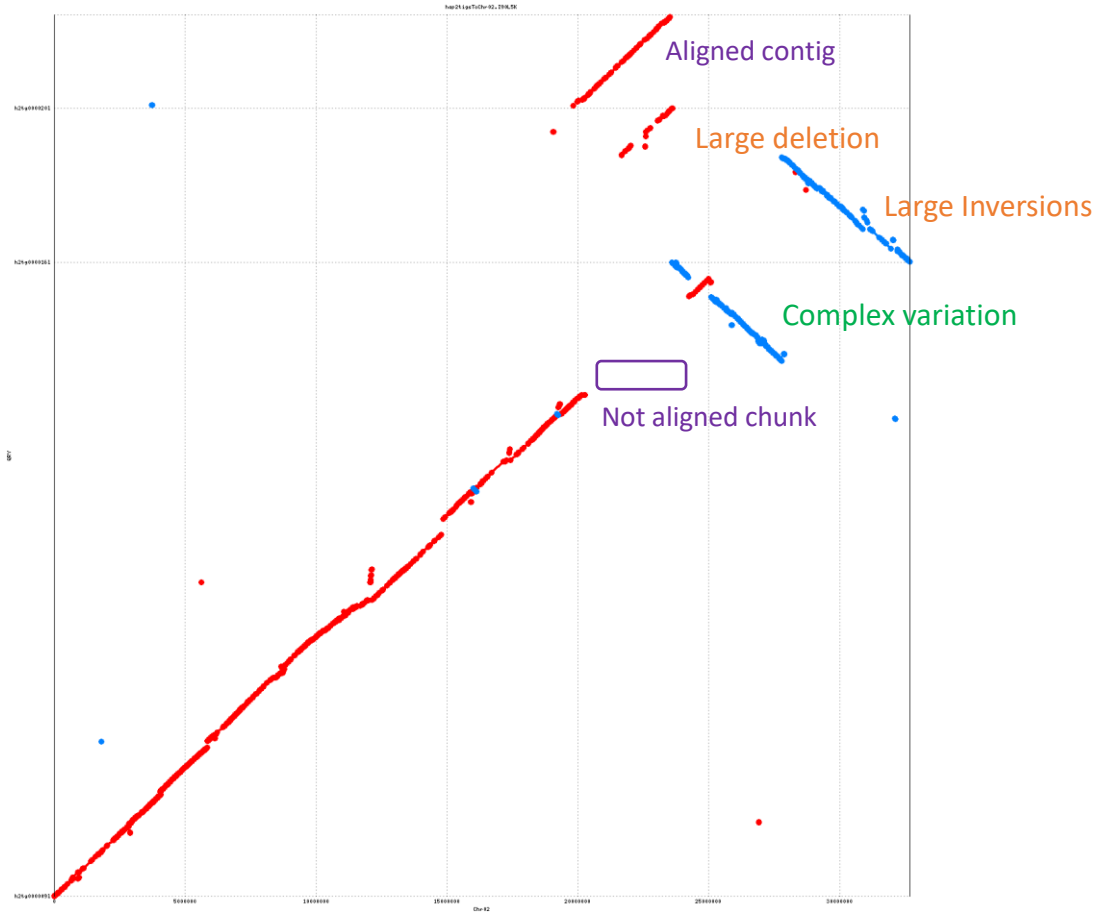
Hap2

Original Chr07_hap2 was super short (242kb) and rebuilt

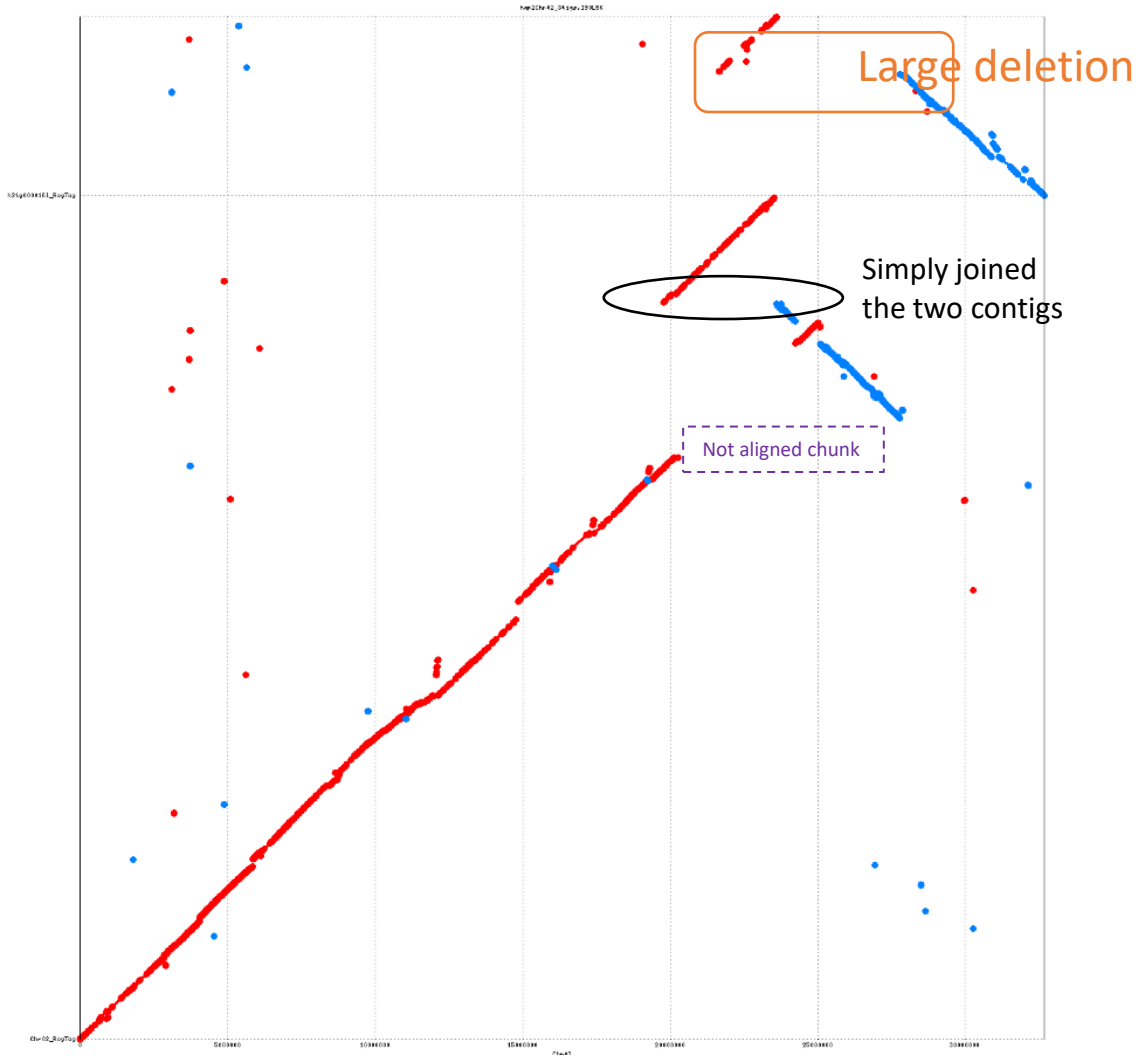
Chr02 joined
NCBI chr02 and chr07

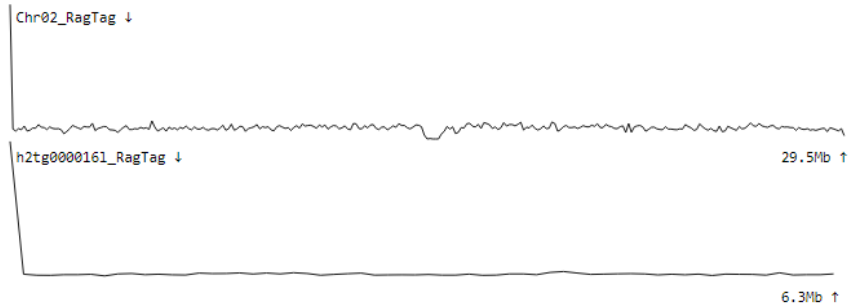


Hap2: Chr07 is fixed. Chr02 is not easy



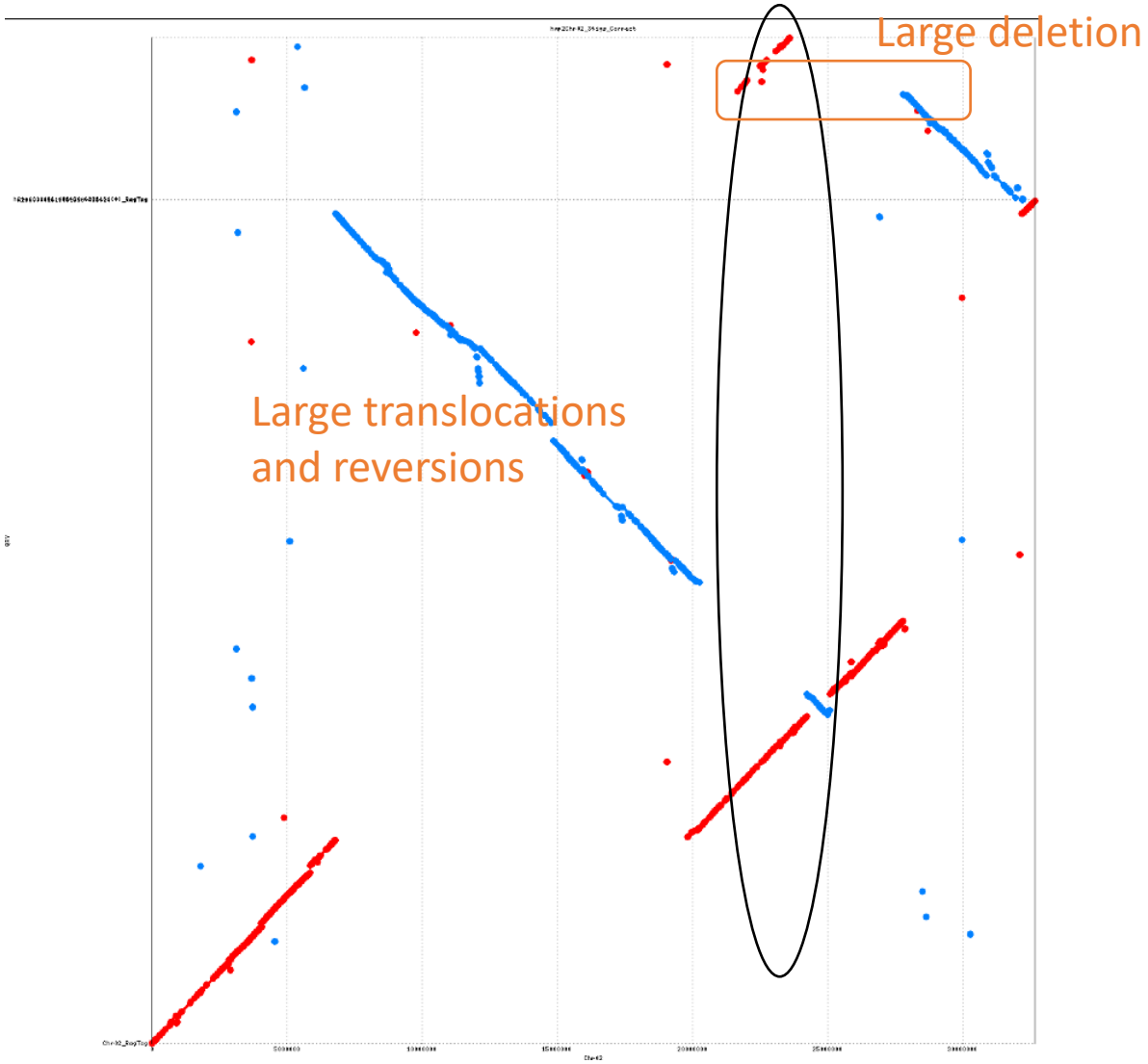
Gillenia hap2: Contigs joined to make chr02





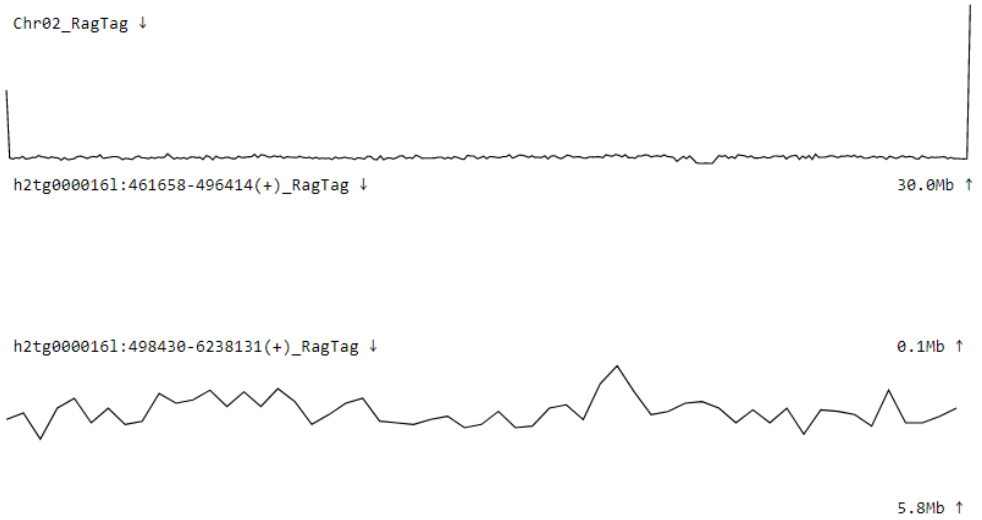
After scaffolding, we got chr02 and the untouched h2tg000016l contig, both with long telomeres at 5’

Gillenia hap2: Contigs corrected and joined to make chr02



Synteny to NCBI Chr02

Telomeres distribution



- Region difficult to solve is circled
- After scaffolding, we got chr02 and two chucks from h2tg000016l
- The new Chr02 is shorter than NCBI chr02;with large reversed translocated chunks; but having long telomeres at both ends

Repeat Detection and Annotation

Plant TE databases available in the public domain:

- Repbase: Curated TE library. Yearly subscription
- PlantRep: TEs detected in 459 plant species using an uniform pipeline
- RepetDB
- nrTEplants: From Ensembl. Computed after combining TREP, SINEbase, Redat, RepetDB, EDTArice, EDTAmaize, SoybaseTE and TAIR10TE

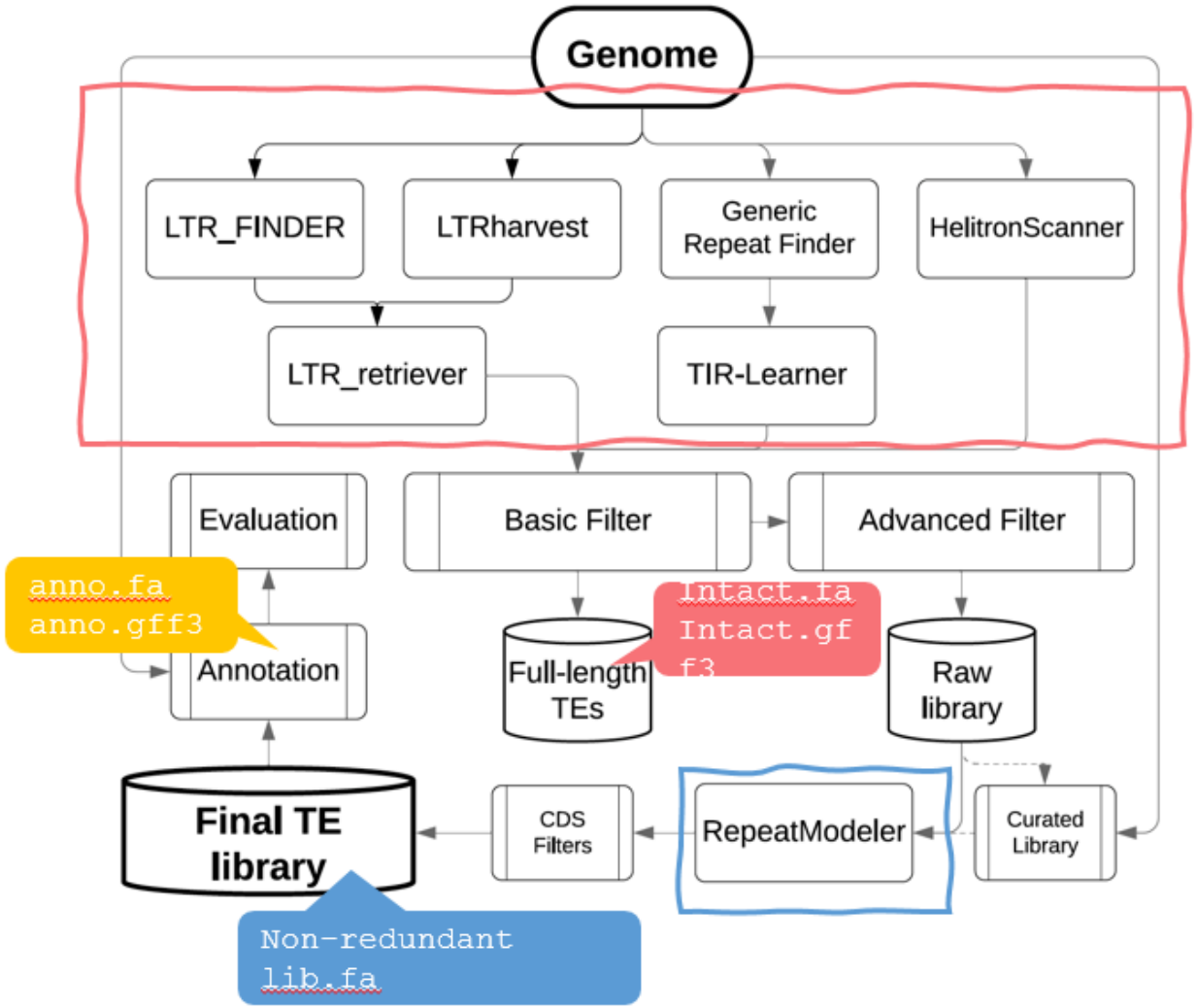
De novo Repeats Discovery Pipelines

- RepeatModeler2 + Dfam or RepBase
- EDTA
- EarlGrey
- REPET (the latest package released in 2019)
- Dfam TETools

De novo Repeats Discovery Pipelines

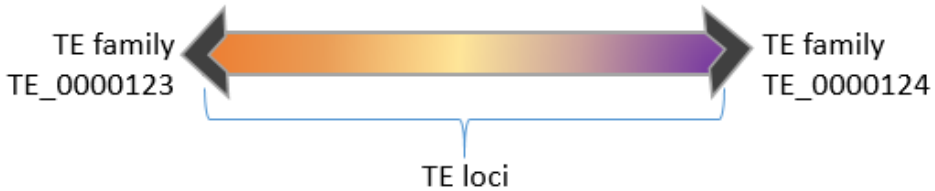
	RepeatModerl 1 + RepBase	EDTA	EarlGrey
Code repo	https://www.repeatmasker.org/RepeatModeler/ https://www.repeatmasker.org/	https://github.com/oushujun/EDTA	https://github.com/TobyBaril/EarlGrey
Pros	Active development Long history Utilized in the other pipelines	Active development Releases: 13 Easy installation since v2 Fast response to questions raised in Github Issues Paper and algorithm published Clear users manual Flexible (re-start from a specified step) Accept --cds (fasta), --curatedlib (fasta) and --rmout (your homology based TE annotation)	Releases: 3 Visualized summary Accept -l == Repbase species subset library (FASTA format) RepeatCraft for de-fragmentation and annotation
Cons	RepBase costs $$ No other TE tools was used	Optional --evaluate step is slow (evaluate classification consistency of the TE annotation) Optional --sensitive mode is slow	Lack of documentation on the actual workflow Lack of document on result interpretation No issue on github so far Very difficult to install

An overview of EDTA pipeline



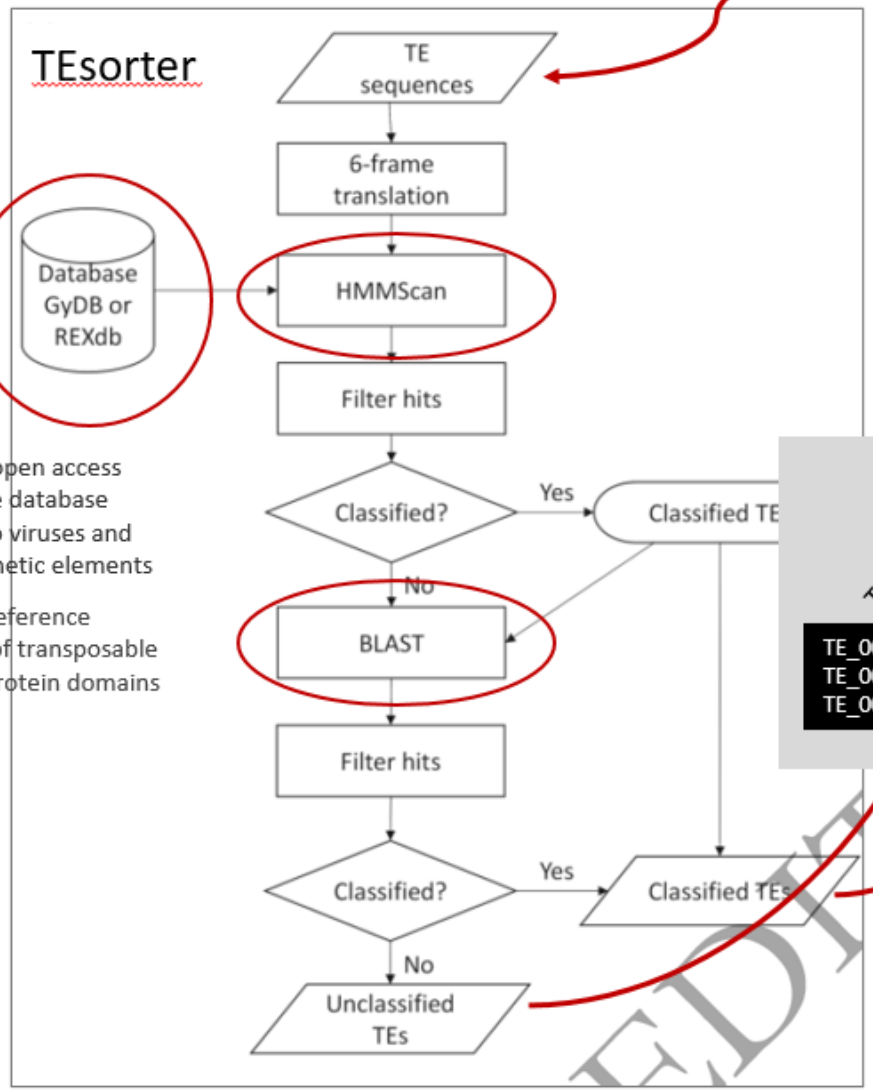
Ou et al. 2019. *Genome Biology*

- **EDTA**: Extensive de-novo TE Annotator
- Identification & classification
- **Identification**: Structural-based + homology-based methods
- **Classification**:
  - Class > Order > Superfamily > Clade > Family
  - Spectrum of similarity



  - Structural characteristic:
    - Target site duplication
    - Terminal motifs
    - Conserved protein domains etc.
  - The 80-95-80 rule (identity-coverage-length)
- **Representative** seq of TE family: canonical vs random pick
- **False positives**: Nested TE/gene, SSR etc.
- **Pan-genome annotation:**
  [separate -> unite] vs [unite -> separate]

Refining TE annotation: 4. annotate conserved protein domains and add labels



GyDB: an open access knowledge database devoted to viruses and mobile genetic elements

REXdb: a reference database of transposable element protein domains

Superfamily
- 3.01_Copia
- 3.02_Gypsy
- 3.03_LTR_Unknown
- 3.04_LINE_Unknown
- 3.05_CACTA
- 3.06_hAT
- 3.07_Mutator
- 3.08_PIF_Harbinger
- 3.09_Tc1_Mariner
- 3.10_Helitron

Check classification consistency:
EDTA > TEsorter

Compile classification

TE_family	Clade	Superfamily	Order	Class	Complete_dom	Strand_dom	Domains	Complete_struc
TE_00001233	Reina	Gypsy	LTR	Retrotransposon	no	+	PRO|Reina	no
TE_00001234	Tekay	Gypsy	LTR	Retrotransposon	yes	+	GAG|Tekay PROT|Tekay RT|Tekay RH|Tekay INT|Tekay	yes
TE_00001235	unknown	hAT	TIR	DNA_transposon	unknown	+	Tpase|hAT	yes

Zhang et al. 2022. *Horticulture Research*

Plan for TE annotation

- EDTA 1st run
- Refine 1st EDTA TE annotation
- Establish curated TE lib
  - High quality intact TE seq (LTR-TE, TIR-TE, Helitron) from 1st TE annotation
  - High quality intact TE seq (LTR-TE, LINE, SINE, TIR-TE, Helitron) from Repbase (e.g. rice, maize, Arabidopsis)
  - High quality intact TE seq from nrTEplants (e.g. those TEs with full-set protein conserved domains annotated by TEsorter)
- EDTA 2nd run with the supplement of the curated TE lib