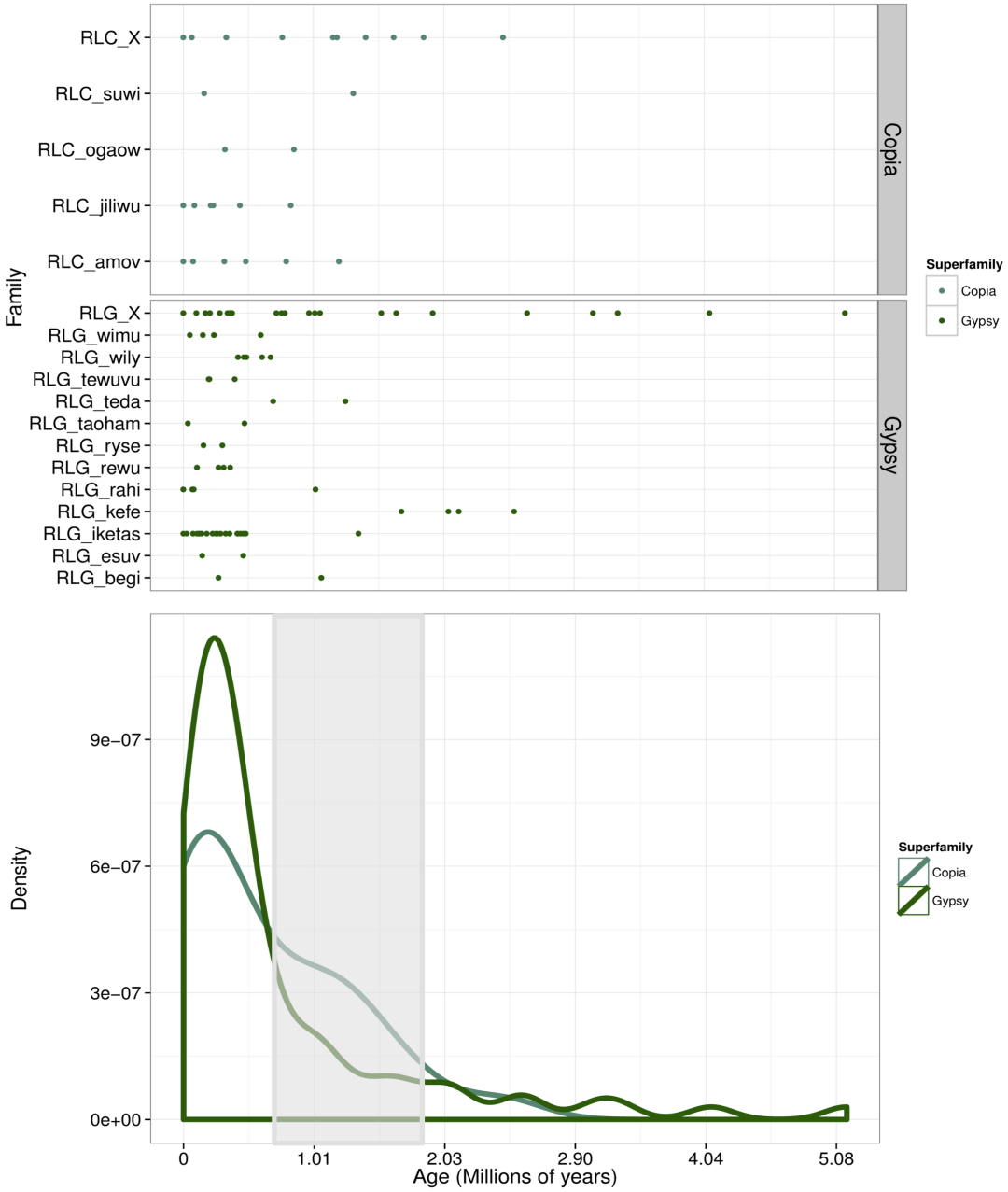
Shared patterns of repeat composition in *Helianthus* provide insight into the history of sunflower genomes

Evan Staton
Department of Botany and Biodiversity Research Centre, UBC
sunflower meeting, 1/14/15

Common sunflower genome description

- Fine-scale structure of the TEs in the sunflower genome has been thoroughly analyzed
  - Time scales of activity
  - Deletion rates
  - Characteristics of insertion to describe genomic bias
  - Differential gain/loss of TEs over time

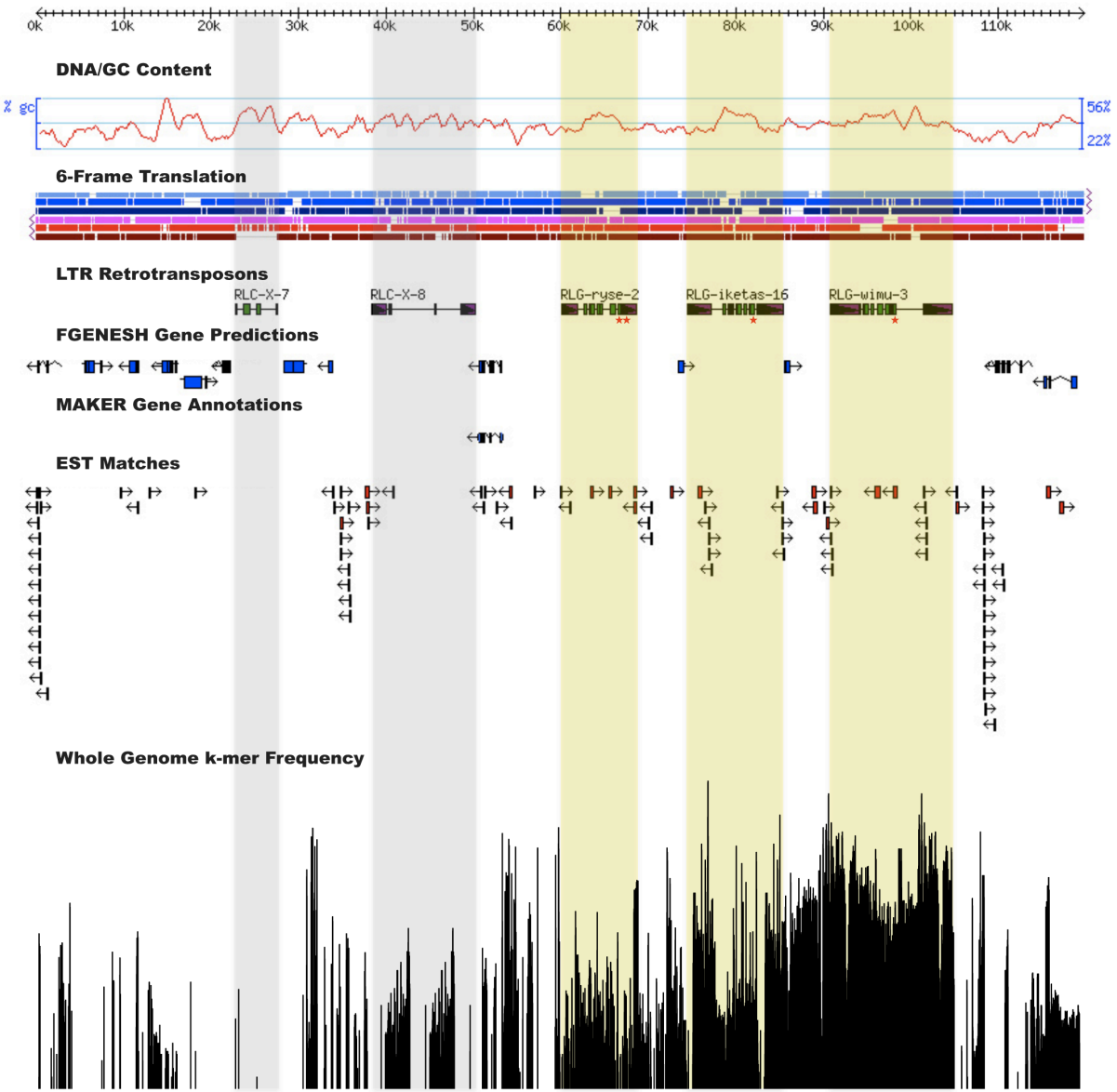
What is the time scale over which sunflower LTR elements have been active?



Est. origin of *H. annuus* (Strasburg and Rieseberg 2008)

Fine-scale structure of a *H. annuus* BAC clone

Gold: Chromoviruses
Grey: *Copia*



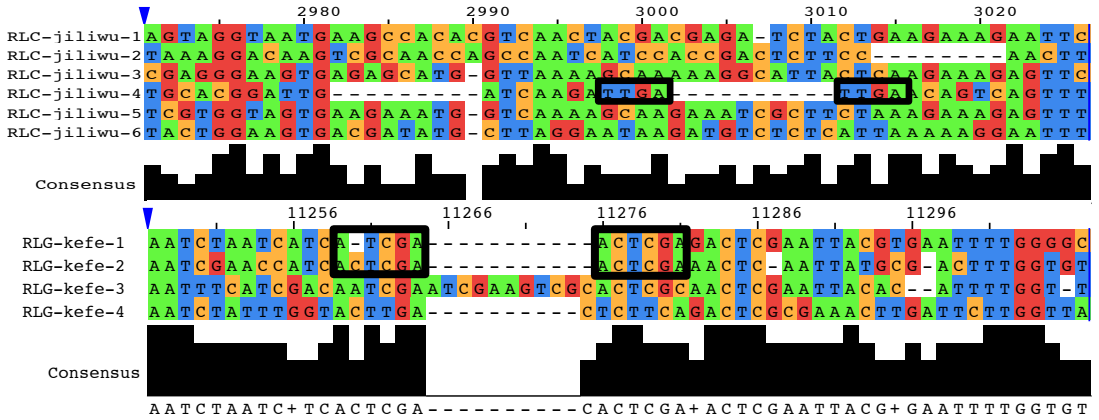
Staton *et al.* 2012, TPJ

Inferring recombination events

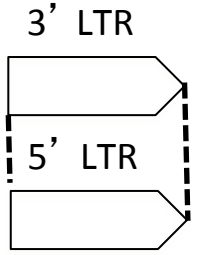
Is there any pattern with respect to element loss?

0.14 : 1.0 : 0.6
solo : intact : truncated

0.2 : 1.0 maize



1. Global alignment
2. Construct global HMM (only in HMMER2)
3. Search BACs/Scaffolds/etc. that have been masked with the fl LTR-RTs



DNA removal in the *H. annuus* genome

Superfamily	Count	Overall length[1]	LTR length[1]	Percent of BACs[2]	Solo:FL:TR[3]
Copia	*28*	*9061*	*775*	*9.86 ± 10.6*	*0.53:1:0.03*
Gypsy	*79*	*9918*	*1551*	*30.47 ± 26.7*	*0.15:1:0.07*
Total	*107*	*9693*	*1346*	*40.33 ± 24.0*	*0.14:1:0.06*

Superfamily	Percent of WGS reads[2]	LTR:RVT[4]
Copia	*19.83 ± 2.8*	*2.27:1*
Gypsy	*57.93 ± 1.4*	*1.53:1*
Total	*77.75 ± 1.84*	*1.9:1*

1 – Lengths are presented as the average (in bp).
2 – Percent composition of BAC clones and WGS reads along with the standard deviation for each superfamily.
3 – Ratio of solo LTRs (Solo) to full-length (FL) to truncated (TR) LTR retrotransposon copies.
4 – The ratio of BLAST hits for LTR sequences (LTR) to reverse transcriptase (RVT) sequences from the WGS reads.

Staton *et al.* 2012, TPJ

Primary research questions

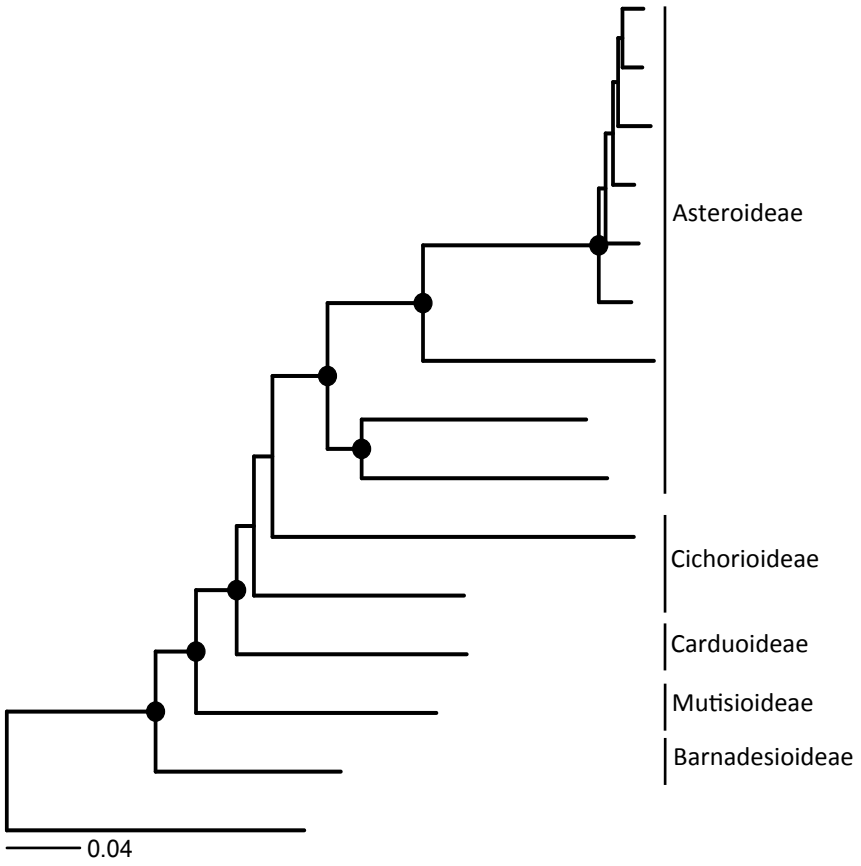
I. How do the patterns in *H. annuus* compare with other Asteraceae
   I. Do we see predominant pattern of *Gypsy* bias?
   II. Is there any phylogenetic signal in TE activity?

Primary research questions

I. How do the patterns in *H. annuus* compare with other Asteraceae
   I. Do we see predominant pattern of *Gypsy* bias?
   II. Is there any phylogenetic signal in TE activity?

II. What level of genome variation is there across all breeding lines of *H. annuus*?

Taxon sampling for WGS sequencing



Subfamily	Species
Asteroideae	***Helianthus argophyllus***
	Helianthus annuus
	Helianthus niveus ssp. tephrodes
	Helianthus verticillatus
	Helianthus porteri
	Phoebanthus tenufolius
	Conoclinium coelestinum
	Pseduognaphalium helleri
	Senecio vulgaris
Cichorioideae	*Taraxacum kok-saghyz*
	Centrapalus pauciflorus
Carduoideae	*Carthamus tinctorius*
Mutisioideae	*Gerbera hybrida*
Barnadesioideae	*Fulcaldea stuessyi*
	Nasanthus patagonicus

Sequencing

Every major subfamily represented

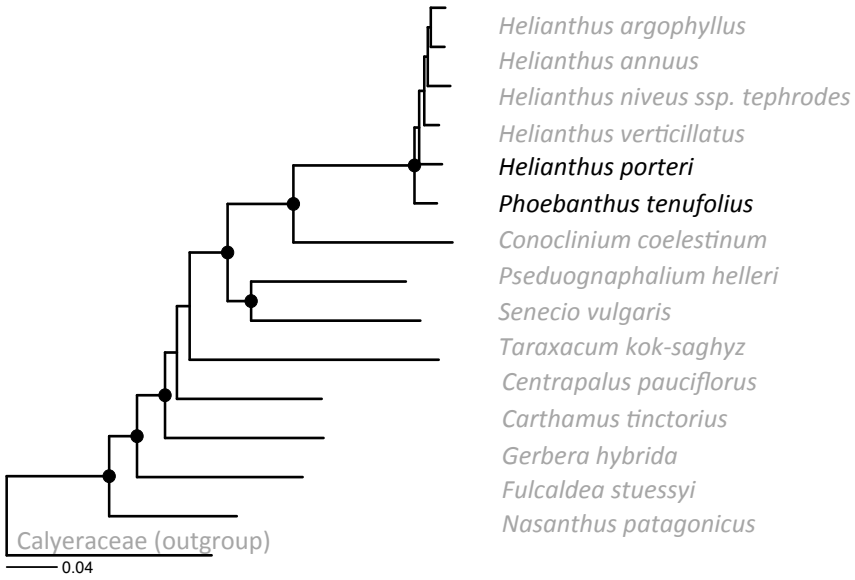
Six species in the tribe Heliantheae

1-5X 100 bp Illumina paired-end WGS

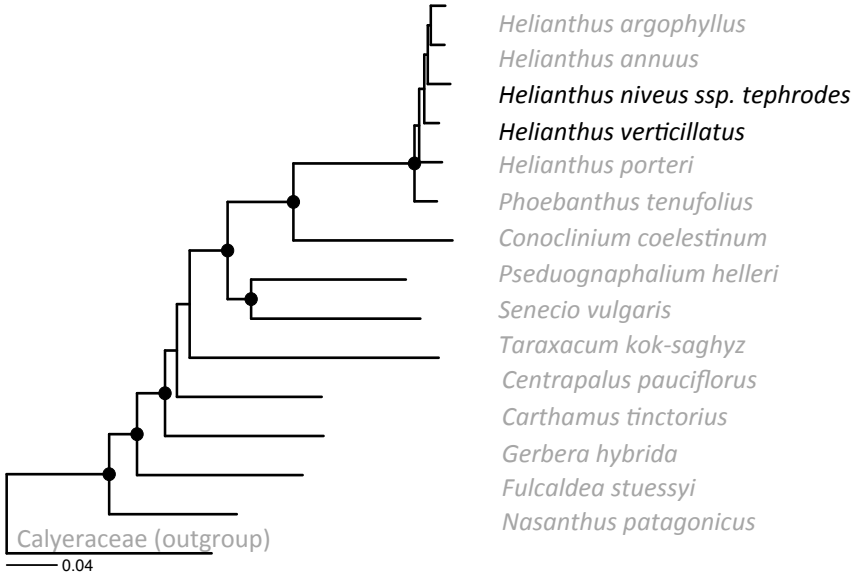
763 nuclear loci targeted by sequence capture

● >75% bootstrap support

Taxon sampling for WGS sequencing

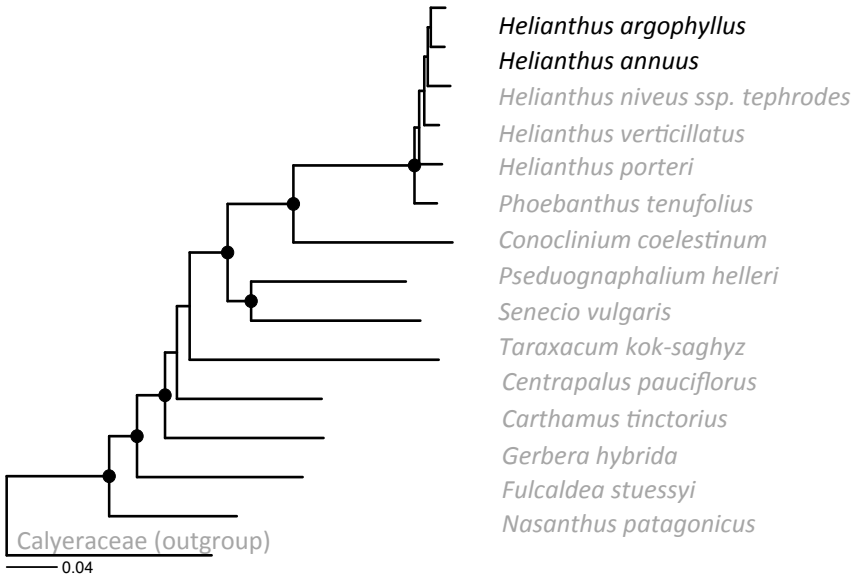


Taxon sampling for WGS sequencing



Taxon sampling for WGS sequencing

Photos courtesy Nolan Kane



Determining the community structure of repeats

- WGS reads offer a non-biased sample of the genome
- Increased data size means greater computational demands -- performance that scales with modern data (i.e., Illumina)

Staton and Burke 2015, *Bioinformatics*
sestaton.github.io/Transposome

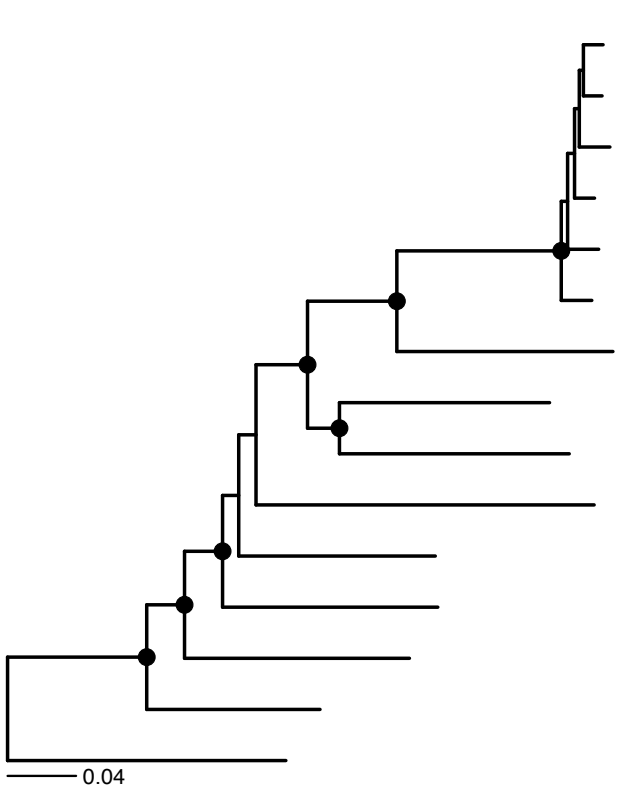
Determining the community structure of repeats

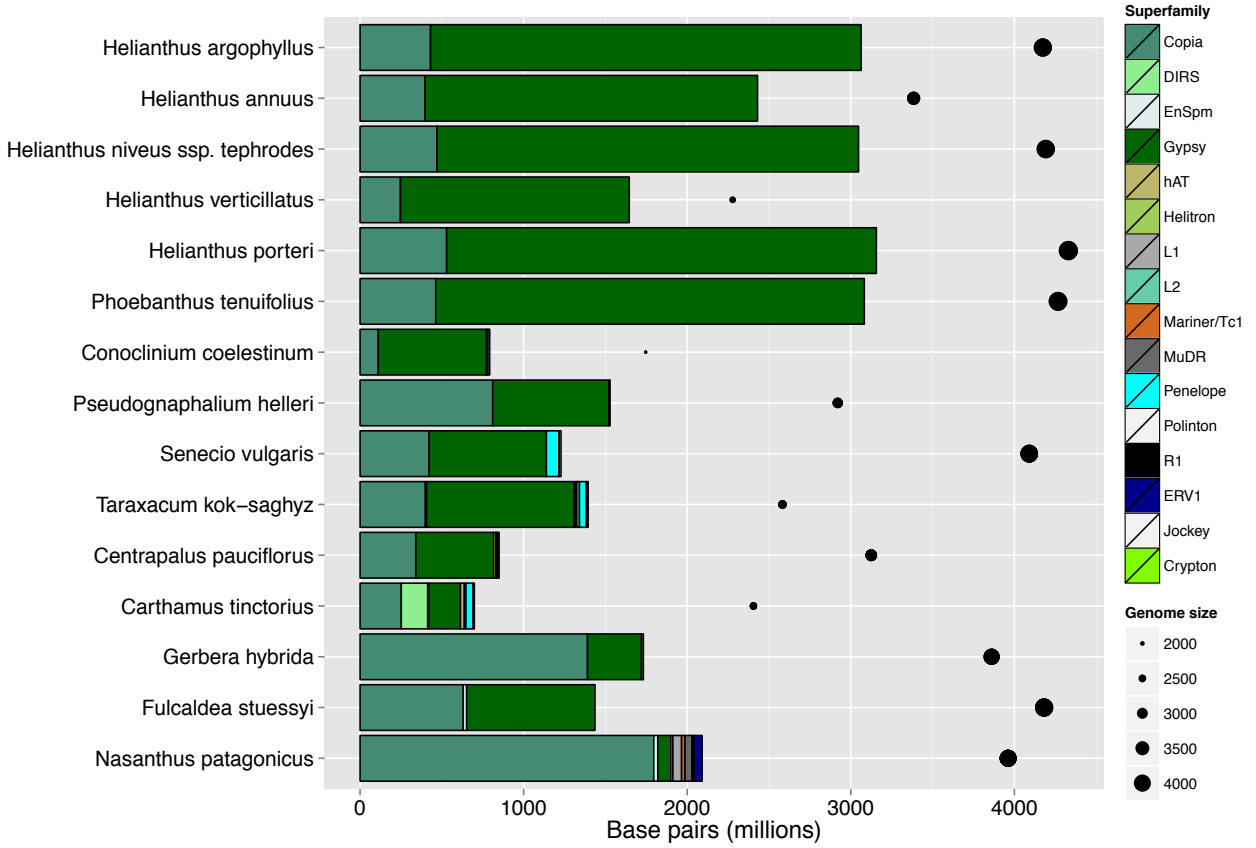
- WGS reads offer a non-biased sample of the genome
- Increased data size means greater computational demands -- performance that scales with modern data (i.e., Illumina)

1) Highly parallel BLAST to determine graph edges
2) Graph-based clustering with edge weights using Louvain method
3) Use paired-end information to merge clusters
4) Annotation to the family-level using defined repeat ontology

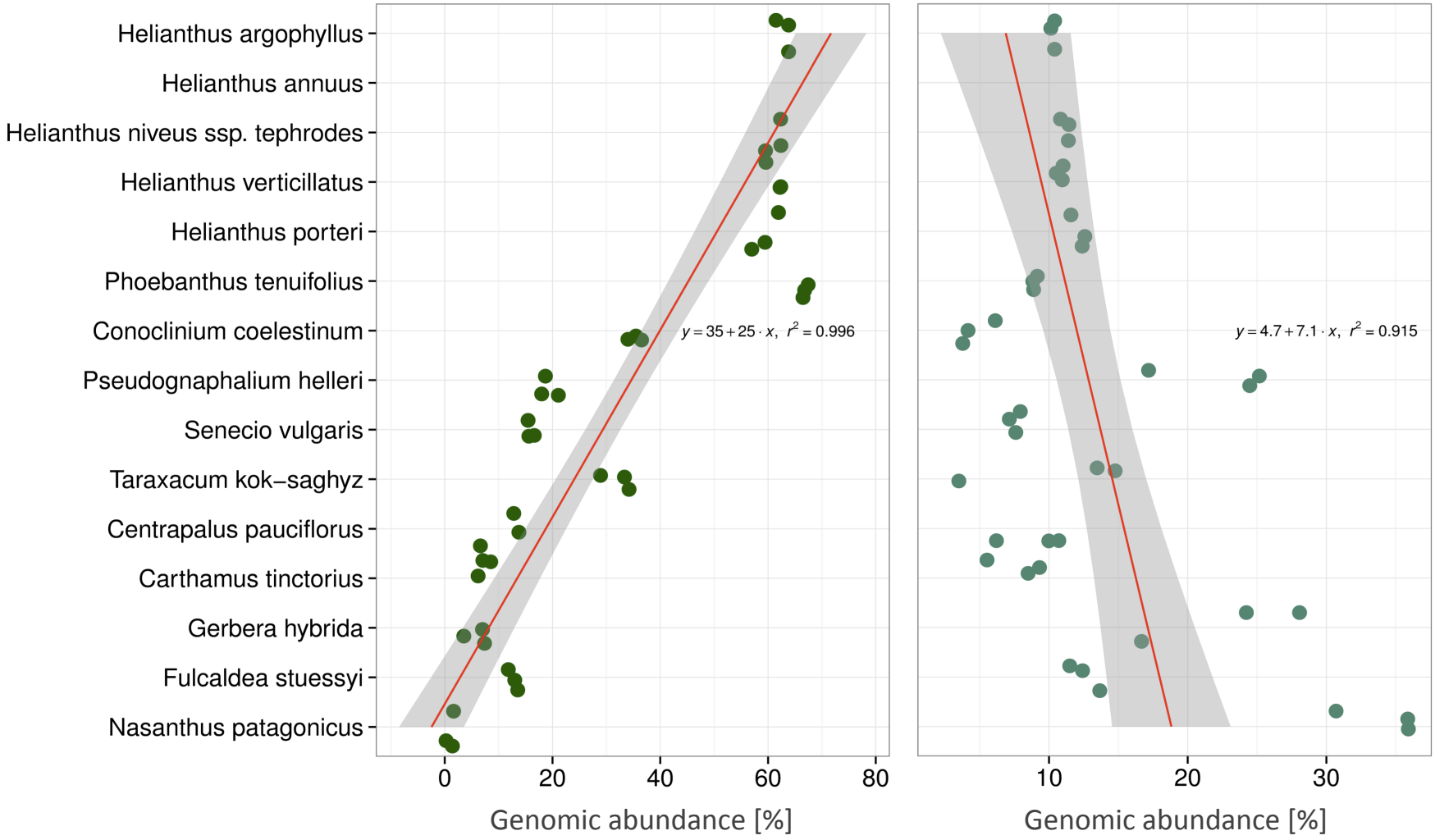
Staton and Burke 2015, *Bioinformatics*
sestaton.github.io/Transposome

Major transitions in genome composition in the Asteraceae

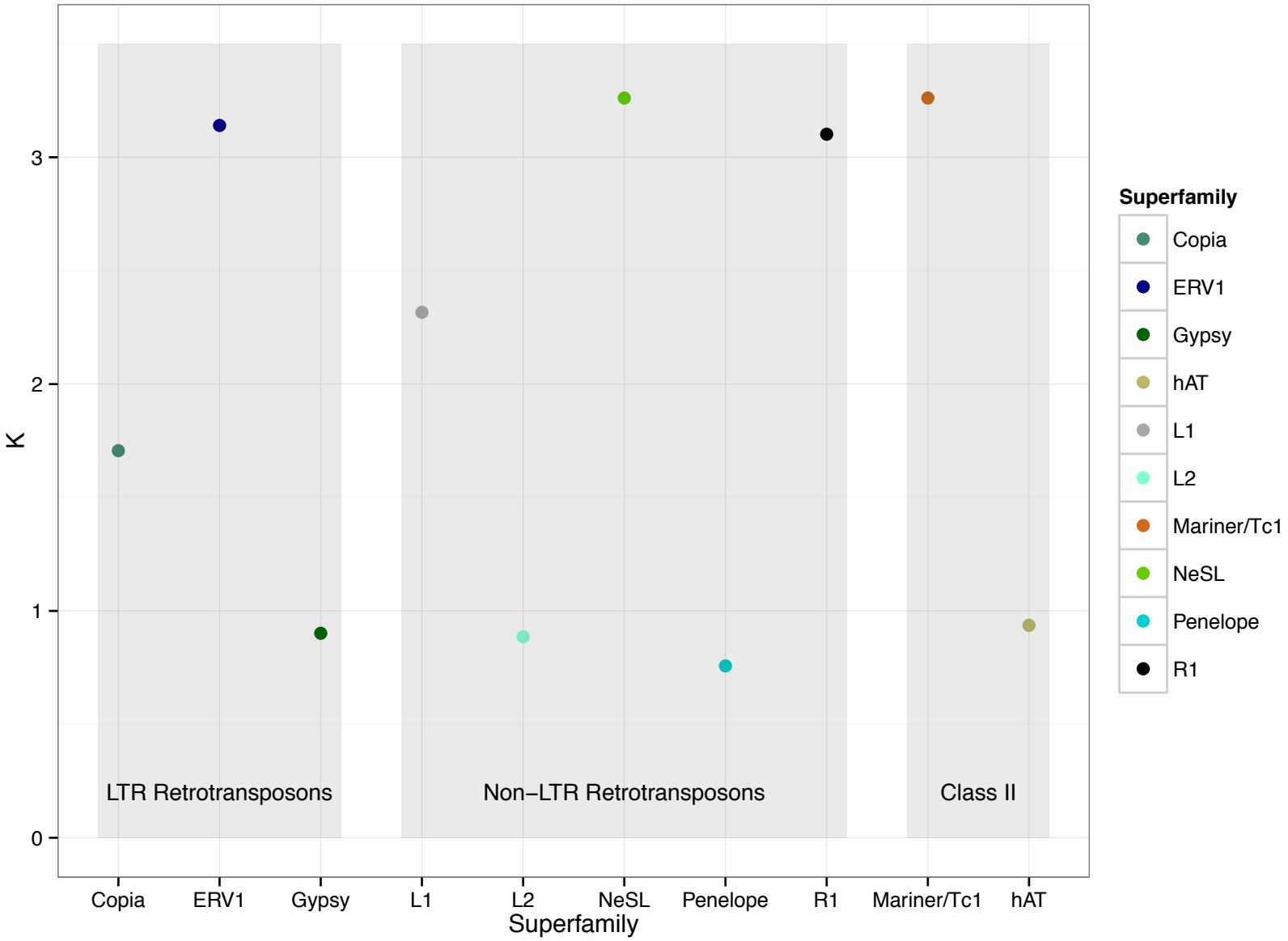




Non-random patterns of change in TE superfamily abundance



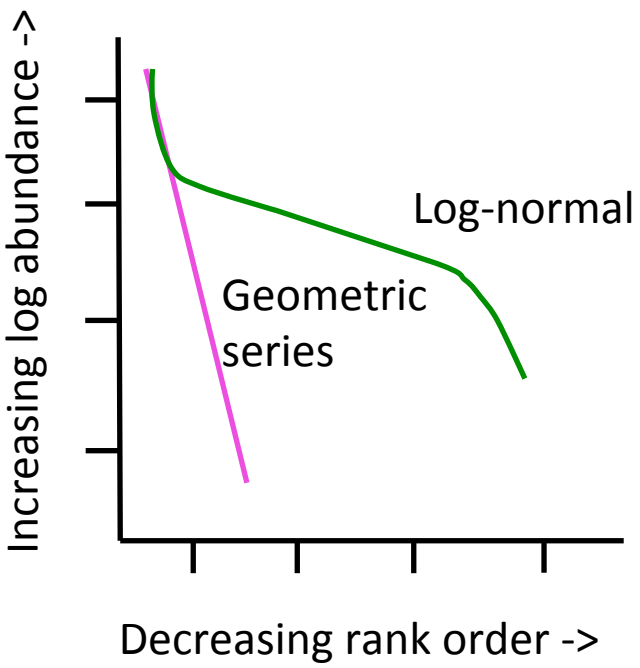
Significant phylogenetic signal for TE superfamilies



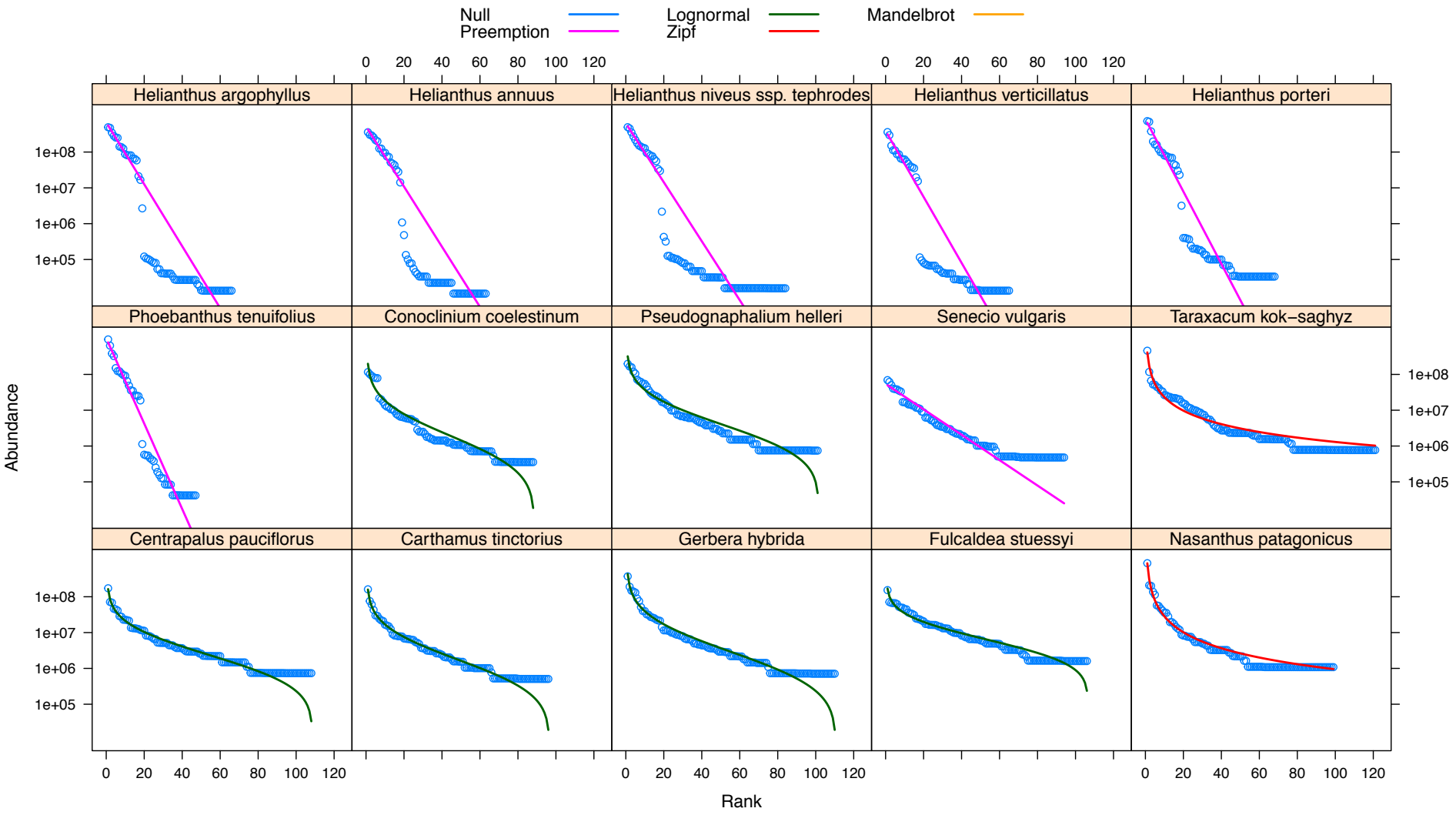
Measuring dynamics with rank abundance/ dominance (RAD) plots

Geometric series (a.k.a. Niche-preemption) – predicts extremely uneven abundances

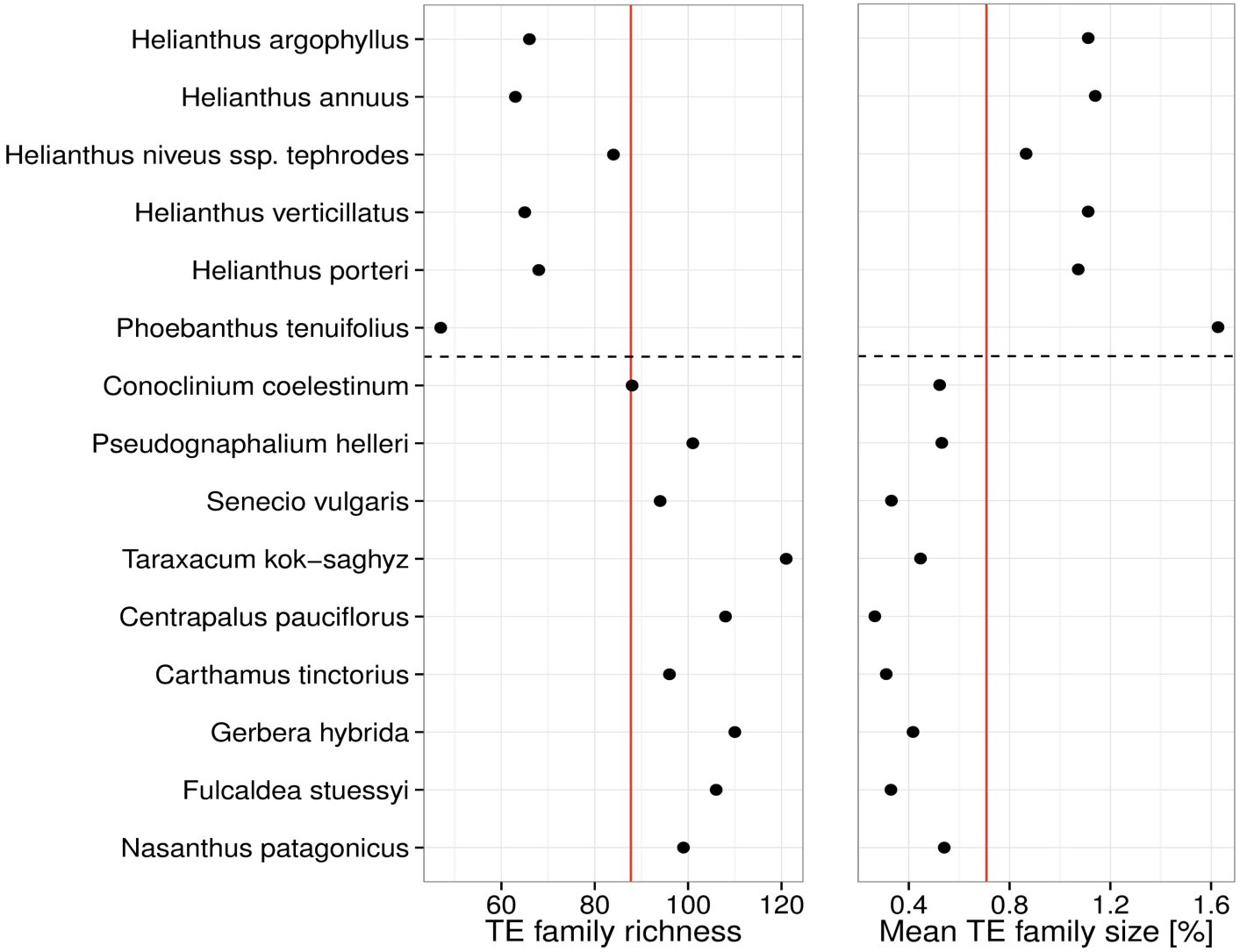
Log-normal – predicts even abundances with low proportion of rare species

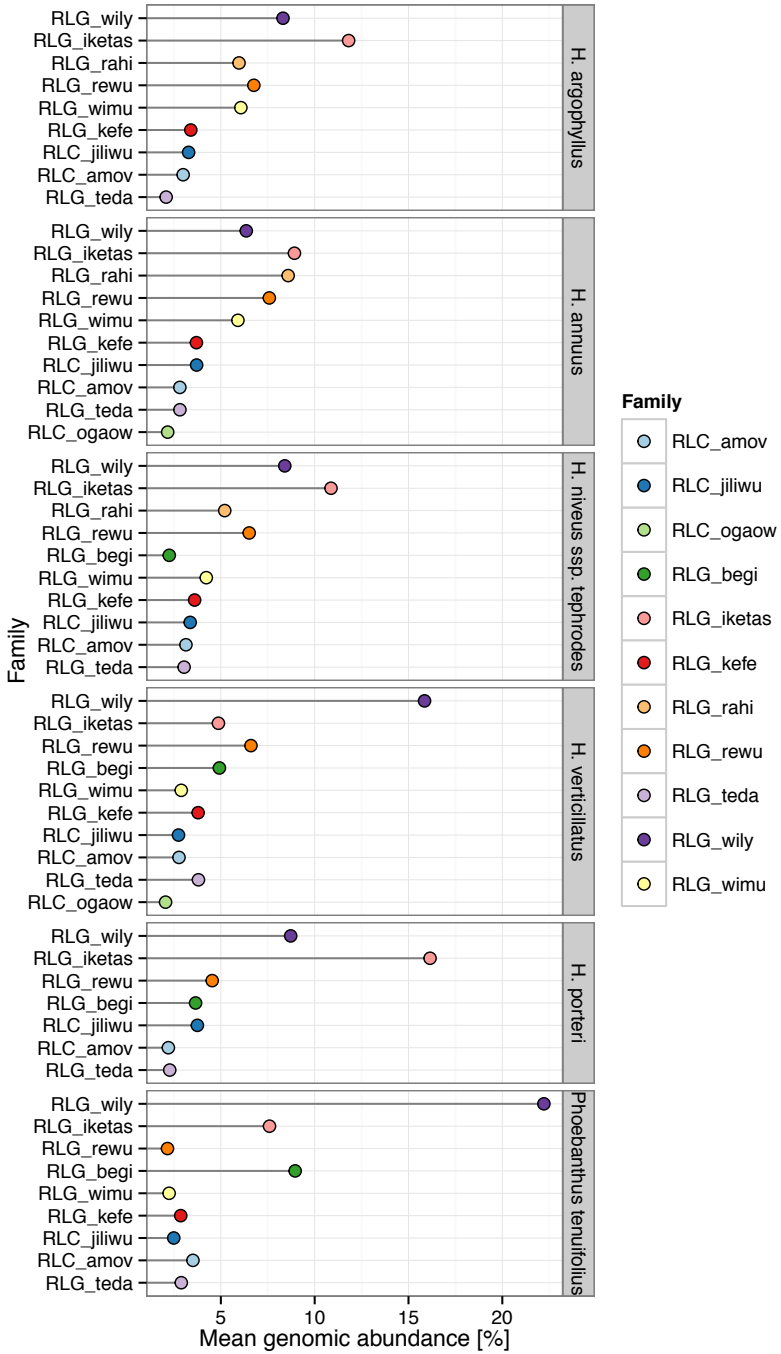


RAD plots for each community of TE families

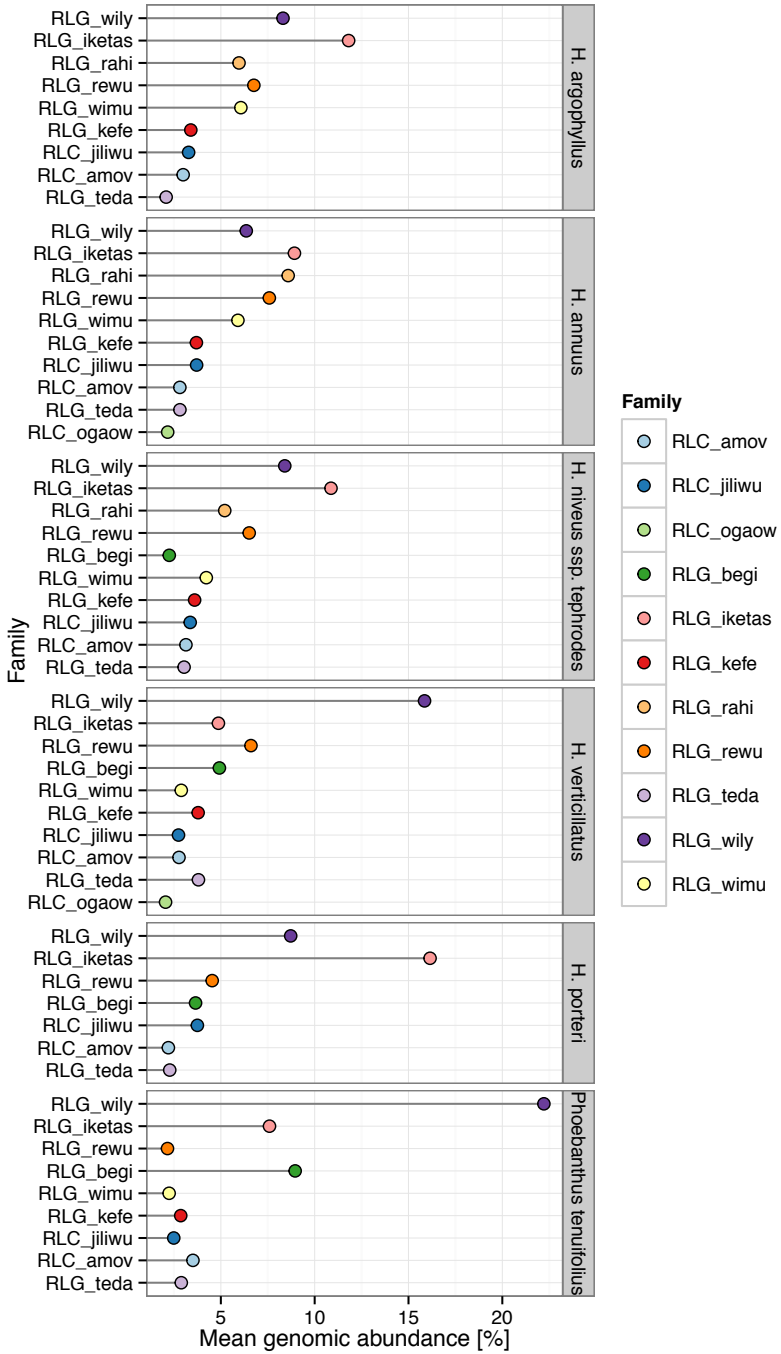


Community level changes in TE properties in the Heliantheae





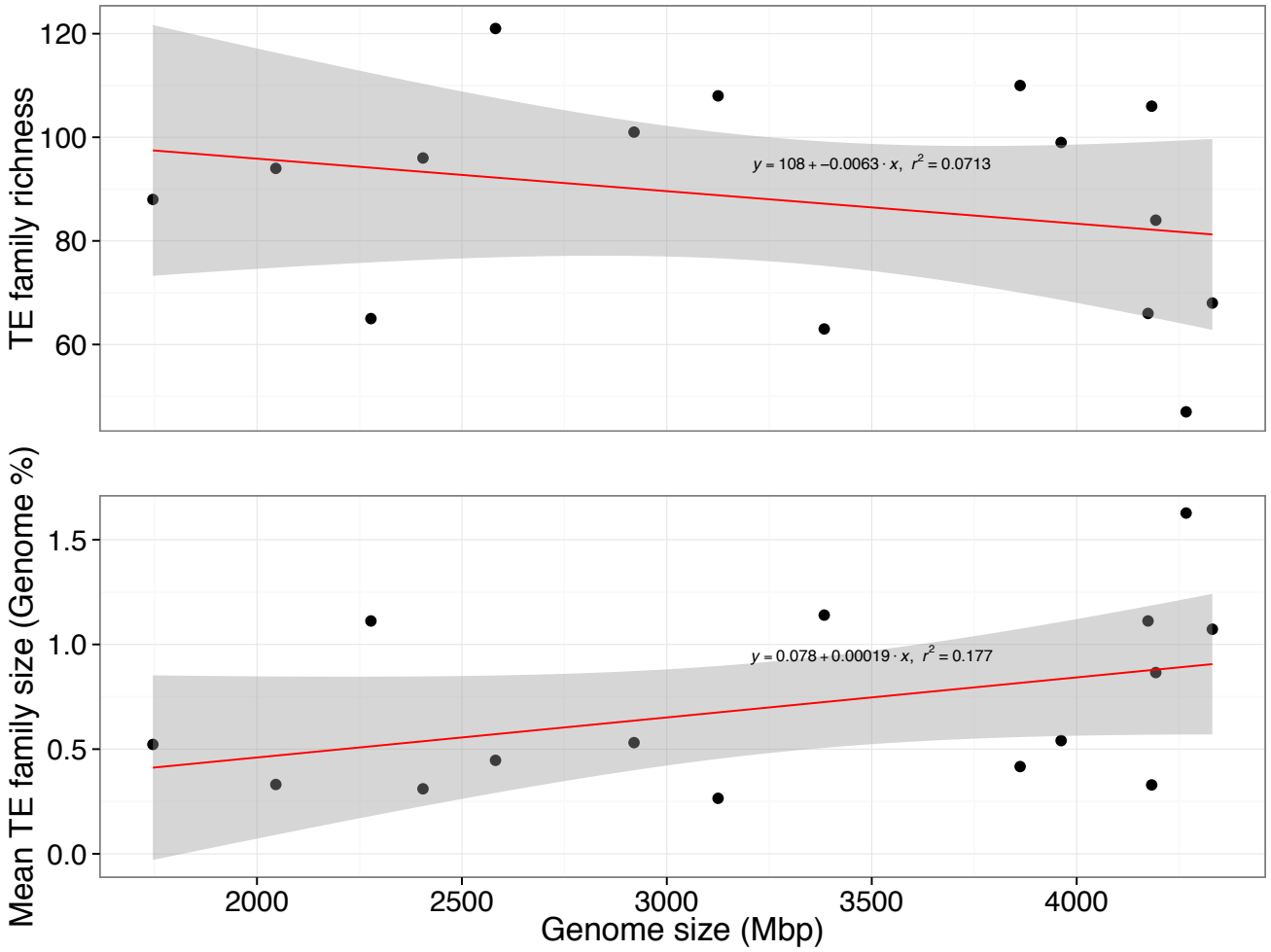
Is there any change in rank order of TE families in the Heliantheae?



Is there any change in rank order of TE families in the Heliantheae?

No phylogenetic pattern of TE family abundance

TE richness is negatively correlated with genome size

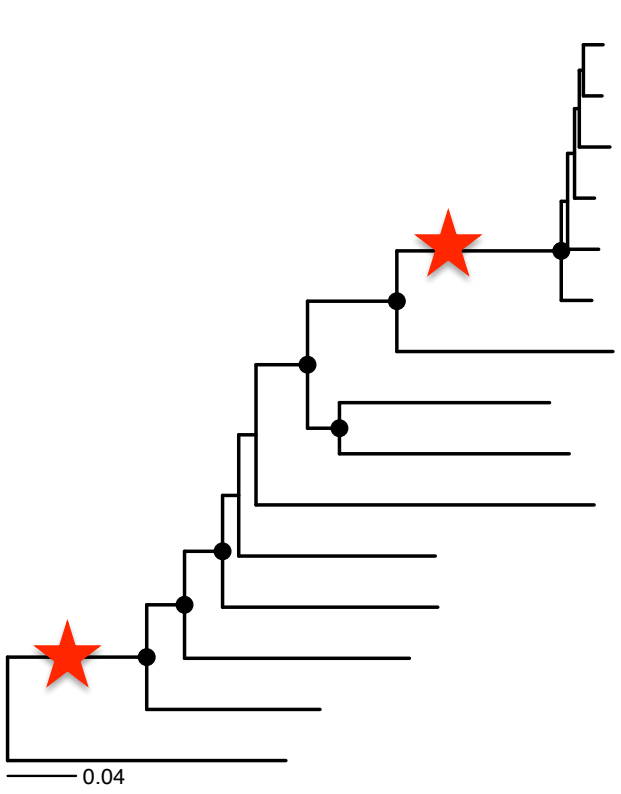


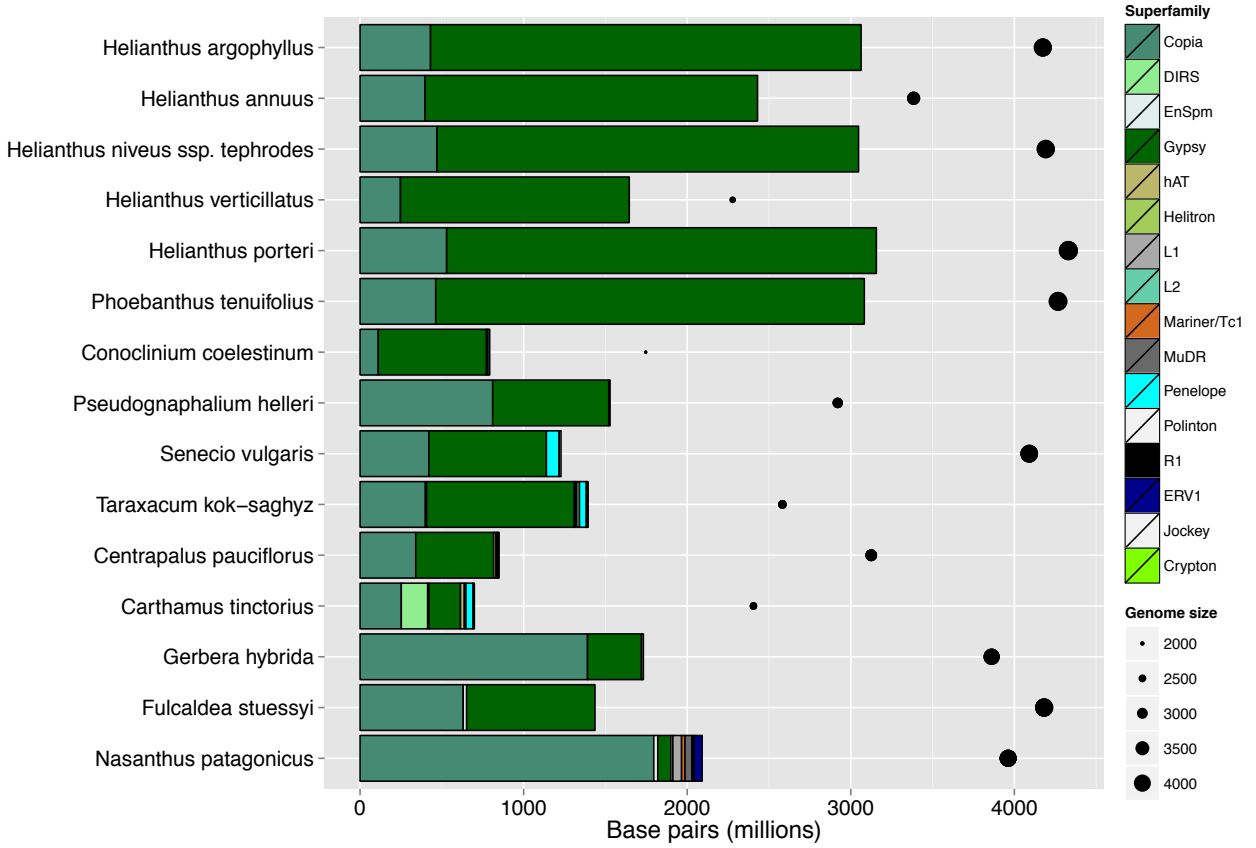
Conclusions

I. How do the patterns in *H. annuus* compare with other Asteraceae
   I. Do we see predominant pattern of *Gypsy* bias?
   II. Is there any phylogenetic signal in TE activity?

- **The genomic properties of *H. annuus* are lineage-specific**
- **Few *Gypsy* families are contributing a large portion of *Helianthus* genomes**

Transitions in genome composition coincide with WGD events





Intraspecific variation in *Helianthus annuus*

Compare 266 sunflower breeding lines representing USDA, INRA, Oil and non-Oil and "core 12" (Mandel et al. 2013)

1) What is the total level of TE variation between lines?

Intraspecific variation in *Helianthus annuus*

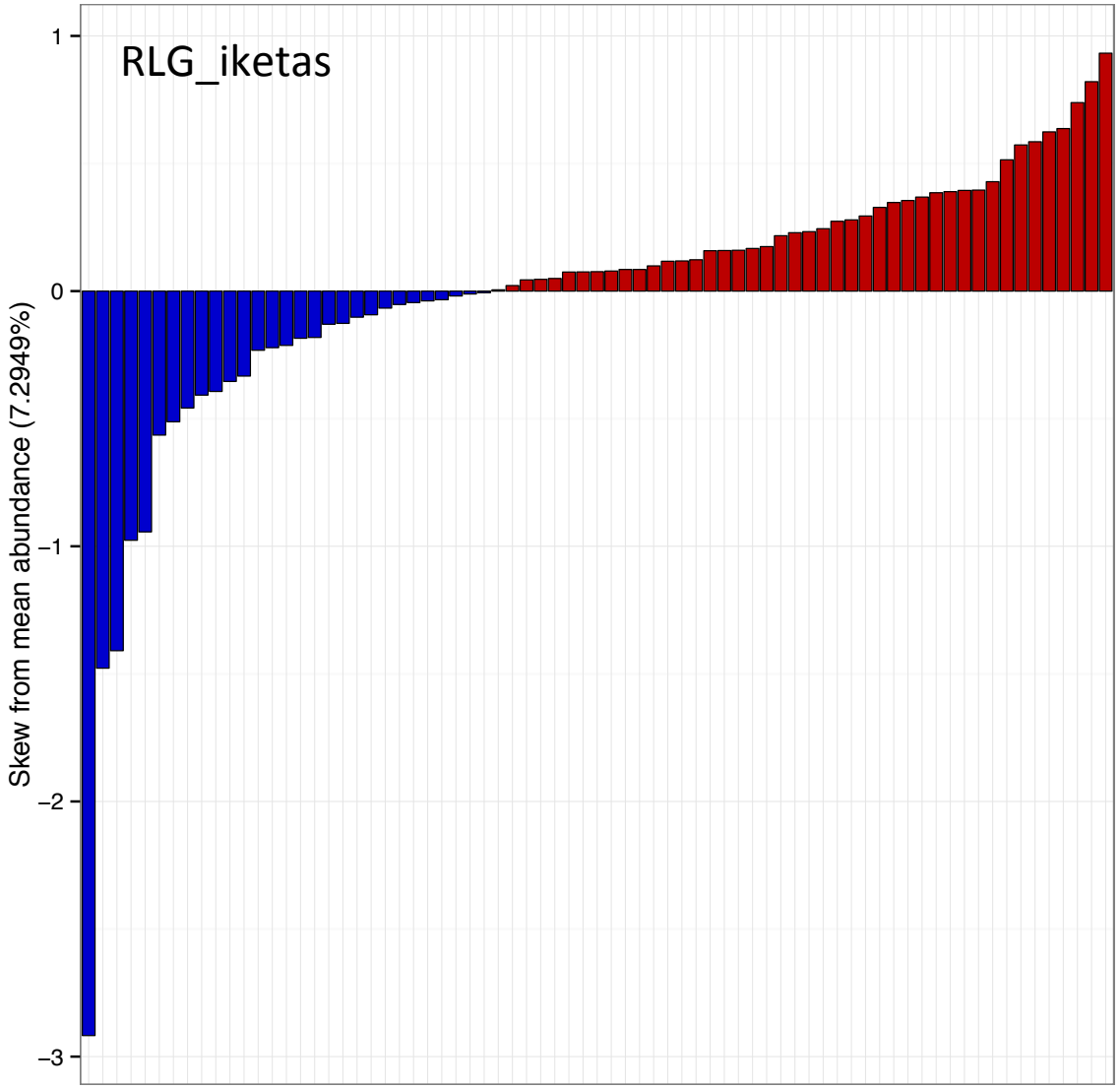
Compare 266 sunflower breeding lines representing USDA, INRA, Oil and non-Oil and "core 12" (Mandel et al. 2013)

1) *What is the total level of TE variation between lines?*

2) *Is there significant variation in TE family abundance between lines?*

3) *Is there any difference between Oil and non-Oil lines?*

Intraspecific variation in *Helianthus annuus*

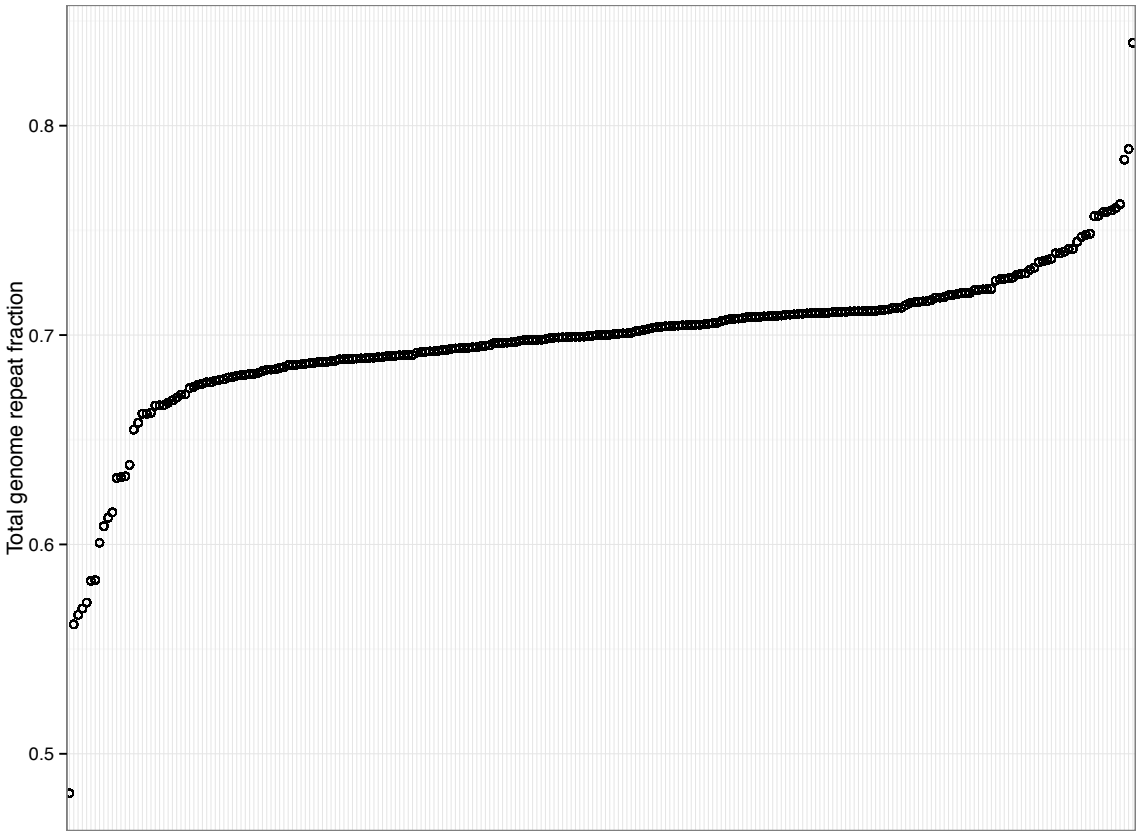


Kruskal-Wallis test for sign. family-level abundance variation

No statistically significant variation of families between lines

Enormous variation exists
*** Two lines differ by 144 Mbp of just one family**

Intraspecific variation in *Helianthus annuus*



Mann-Whitney-Wilcoxon test for sign. family-level variation between Oil and non-Oil lines

Measured the total variation between lines by source

Significant variation between 64 Oil/non-Oil lines

Lines vary by ~30% in total repeat abundance

Min ~48.1% - HA 318
Max ~83.9% - HA 853

Conclusions

Compare 266 sunflower breeding lines representing USDA, INRA, Oil and non-Oil and "core 12" (Mandel et al. 2013)

1) What is the total level of TE variation between lines?

Conservatively, 30% total variation.

Conclusions

Compare 266 sunflower breeding lines representing USDA, INRA, Oil and non-Oil and "core 12" (Mandel et al. 2013)

2) *Is there significant variation in TE family abundance between lines?*

Preliminary analysis suggests no statistical difference, but enormous biological variation exists.

3) *Is there any difference between Oil and non-Oil lines?*

Approx. 25% lines differ significantly.

Acknowledgements

Collaborators

Vicki Funk (Smithsonian)
Mauricio Bonifacino (UR)
Jennifer Mandel (UM)

Staff members

Michael Boyd
- Greenhouse staff

Funding sources

United States Department of Agriculture
National Institute of Food and Agriculture

Special thanks

Donovan lab - UGA

John Burke – UGA
(former lab)
- Lab mates

Rieseberg lab - UBC
(current lab)
- All my current lab mates

GACRC staff

Greg Derda
Yecheng Huang
Shan-Ho Tsai
Paul Brunk (Sys Admin)


Nasanthus patagonicus