Repeat sequence characterization in sugar pine (*Pinus lambertiana*) and loblolly pine (*Pinus taeda*)

Robin Paul[1], Kristian A. Stevens[2], Pedro J Martinez-Garcia [3], Aleksey Zimin[4], Ann Holtz-Morris[5], James A. Yorke[4,6],Maxim Koriabine[5], Marc Crepeau[2], Daniela Puiu[7], Steven L. Salzberg[7], Pieter J. deJong[5], Charles H. Langley[2], Sowmya Kurugunti[1] , David B. Neale[3], Jill L. Wegrzyn[1]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut, USA
[2]Department of Evolution and Ecology, University of California, Davis, Davis, CA
[3]Department of Plant Sciences, University of California Davis, Davis, CA, USA
[4]Institute for Physical Science and Technology, University of Maryland, College Park, MD, USA
[5]Children's Hospital Oakland Research Institute, Oakland, CA
[6]Departments of Mathematics and Physics, University of Maryland, College Park, MD
[7]Johns Hopkins University, School of Medicine, Baltimore, MD

Abstract

The pathogen (*Cronartium ribicola*) which causes white pine blister rust is one of many factors responsible for the decline in the populations of several white pine species, including sugar pine (*Pinus lambertiana*). For conservation of these species and maintenance of genetic diversity, elucidation of the sugar pine genome is essential. The size of conifer genomes range from 6.5-40 gigabases and studies from the first three genomes (Norway spruce, white spruce, and loblolly pine) support that is this is largely due to repetitive sequences. As a result, the assembly and annotation of these complex genomes is challenging. In this study, we have characterized the gene space and repetitive content in the first version of the 34Gb sugar pine genome. A combination of de novo and homology-based methodologies have been employed to comprehensively identify both interspersed and tandem repeat content. Interspersed elements were compared against the plant section of the RepBase database as well as a library of repeats that was constructed de novo by RepeatModeler. The unclassified interspersed elements identified with RepeatModeler were further characterized using a combination of homology and structural-based approaches.

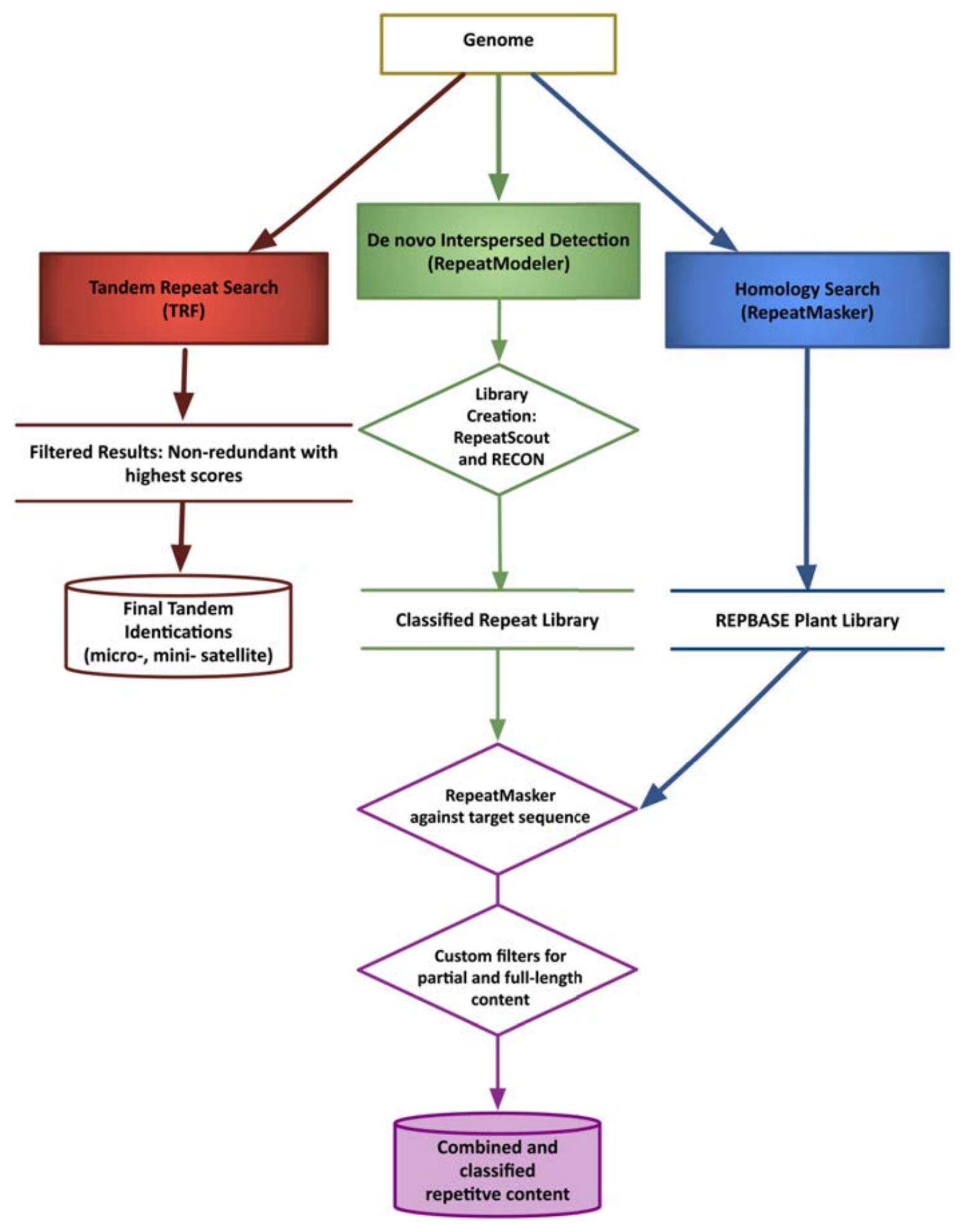
Genome
Tandem Repeat Search (TRF)
De novo Interspersed Detection (RepeatModeler)
Homology Search (RepeatMasker)
Filtered Results: Non-redundant with highest scores
Library Creation: RepeatScout and RECON
Final Tandem Identications (micro-, mini- satellite)
Classified Repeat Library
REPBASE Plant Library
RepeatMasker against target sequence
Custom filters for partial and full-length content
Combined and classified repetitve content

Figure 1: Custom repeat identification pipeline fto combine *de novo* and homology-based approaches for complex non-model genomes

Table 1: Sequence and Repeat Identification Statistics for Conifer Genome and Fosmid Sequences.

	Loblolly pine genome (v1.01)	Sugar pine genome (v0.5) (> 200bp)	Loblolly pine BACs/fosmids	Sugar pine fosmids	Douglas-fir fosmids
No. of scaffolds	1.4×10^7	4.6×10^6	9.1×10^4	5×10^3	3.7×10^3
Total Sequence (bp)	2×10^{10}	2.5×10^{10}	2.7×10^8	1.6×10^8	1.1×10^8
N50 (bp)	5.5×10^4	5×10^4	1.7×10^4	3.3×10^4	3.3×10^4
Percentage of interspersed repeats	72.6	74.8	75.3	75	67
Total full length interspersed repeats (bp)	3.4×10^9	2.4×10^9	4.7×10^7	1.8×10^7	1.3×10^7
Total partial length interspersed repeats (bp)	1.1×10^{10}	1.6×10^{10}	1.6×10^8	1×10^8	6.5×10^7
Percentage of simple repeats	1.8	1.6	1.4	1.4	1.8

Table 2: Comparative repetitive content in sequenced plant genomes.

Species	Genome size (Mb)	Repetitive content (%)
Oryza sativa	362	26
Sorghum bicolor	739	62
Zea mays	2048	85
Glycine max	973	57
Malus x domestica	604	67
Vitis vinifera	477	27
Picea abies	12,019	70
Pinus taeda	22,100	74
Pinus lambertiana	34,000	76

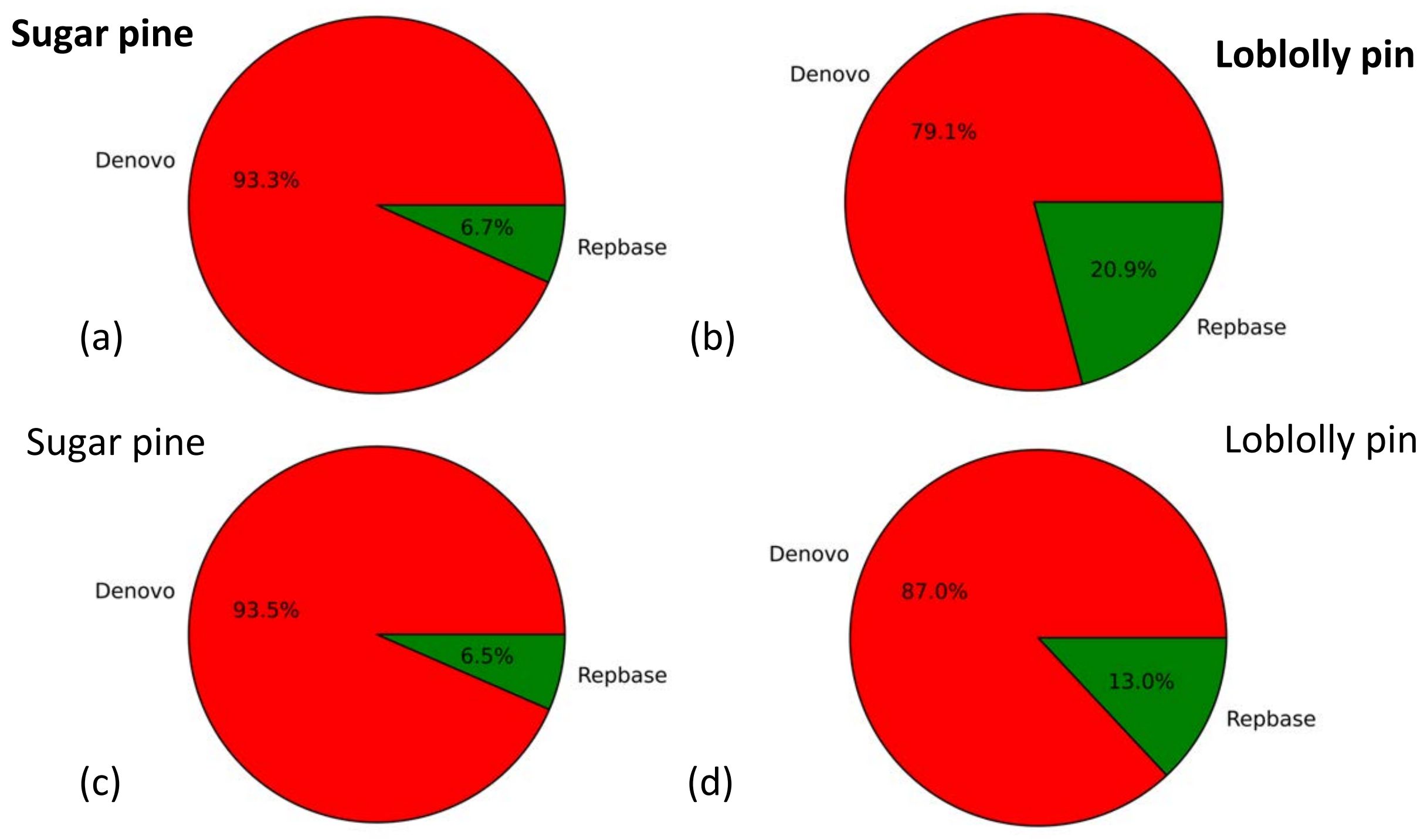

Figure 2: Distribution of full (a & b) and partial (c & d) length alignments between Repeatmodeler and Repbase hits in sugar pine (a & c) and loblolly (b & d).

Table 3: Most prevalent interspersed repeat families in loblolly pine and sugar pine genomes (*de novo* library)

Element name	Source database	Element type	No. of copies	% of genome	Total length (bp)
Loblolly pine					
rnd-2_family-2	RepeatModeler	LTR/Gypsy	524152	5.28	1×10^9
rnd-2_family-115	RepeatModeler	LTR/Copia	944200	2.13	4.3×10^8
rnd-2_family-4	RepeatModeler	LTR/Gypsy	403126	3.03	6.1×10^9
Sugar pine					
rnd-3_family-195	RepeatModeler	LTR/Gypsy	232650	2.03	5.2×10^8
rnd-4_family-638	RepeatModeler	Unknown	276203	1.54	3.9×10^8
rnd-3_family-327	RepeatModeler	LTR/Gypsy	239895	1.49	3.8×10^8

Table 4: Most prevalent interspersed repeat families in loblolly pine and sugar pine genomes (Repbase library)

Element name	Source database	Element type	No. of copies	% of genome	Total length (bp)
Loblolly pine					
PtConagree_I	Repbase	LTR/ Retrotransposons	550351	0.64	$1 .2\times 10^8$
IFG-7a_PTa-I	Repbase	Gypsy-type Retrotransposons	308936	0.35	7.1×10^7
PtCumberland_I	Repbase	LTR/ Retrotransposons	170556	0.3	6.1×10^7
Sugar pine					
PtTalladega_I	Repbase	LTR/ Retrotransposons	3355	0.018	4.5×10^6
PtBastrop_I	Repbase	LTR/ Retrotransposons	2272	0.019	5×10^7
PtAngelina_I	Repbase	LTR/ Retrotransposons	1617	0.012	3.2×10^7

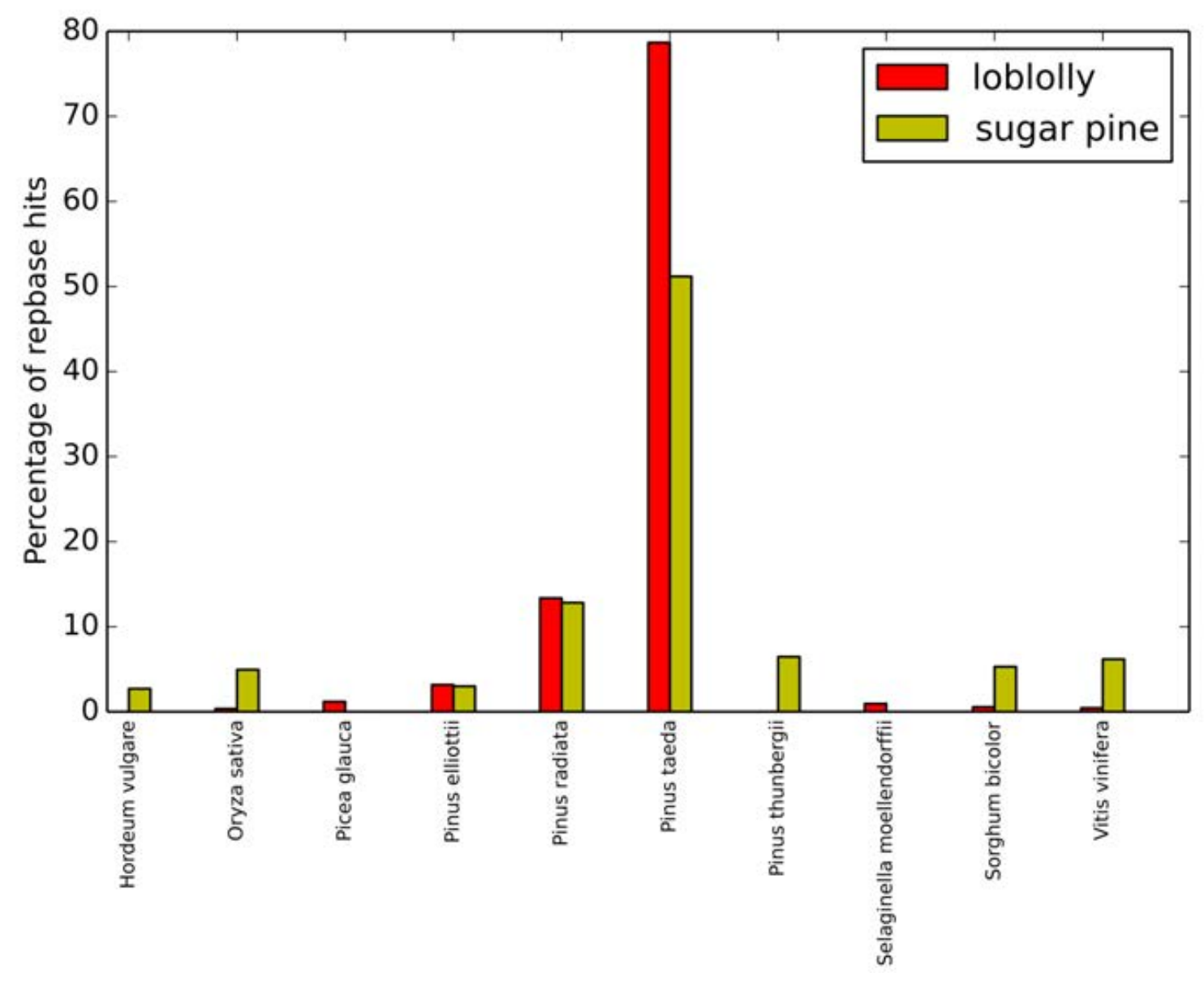

Figure 3: Species origin of plant Repbase matches for loblolly pine and sugar pine

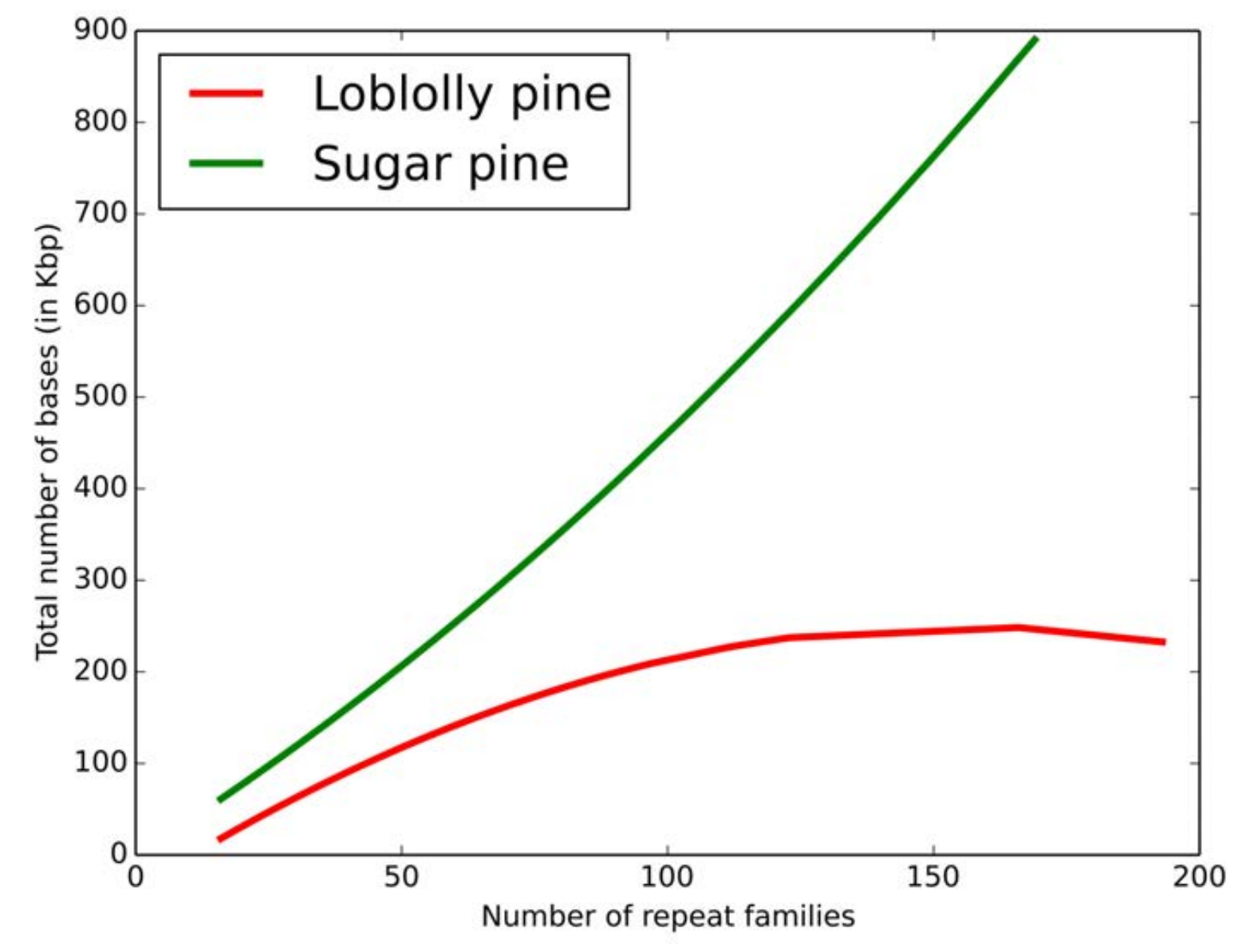

Figure 4: Total repetitive content explained by the top contributing interspersed repeat families

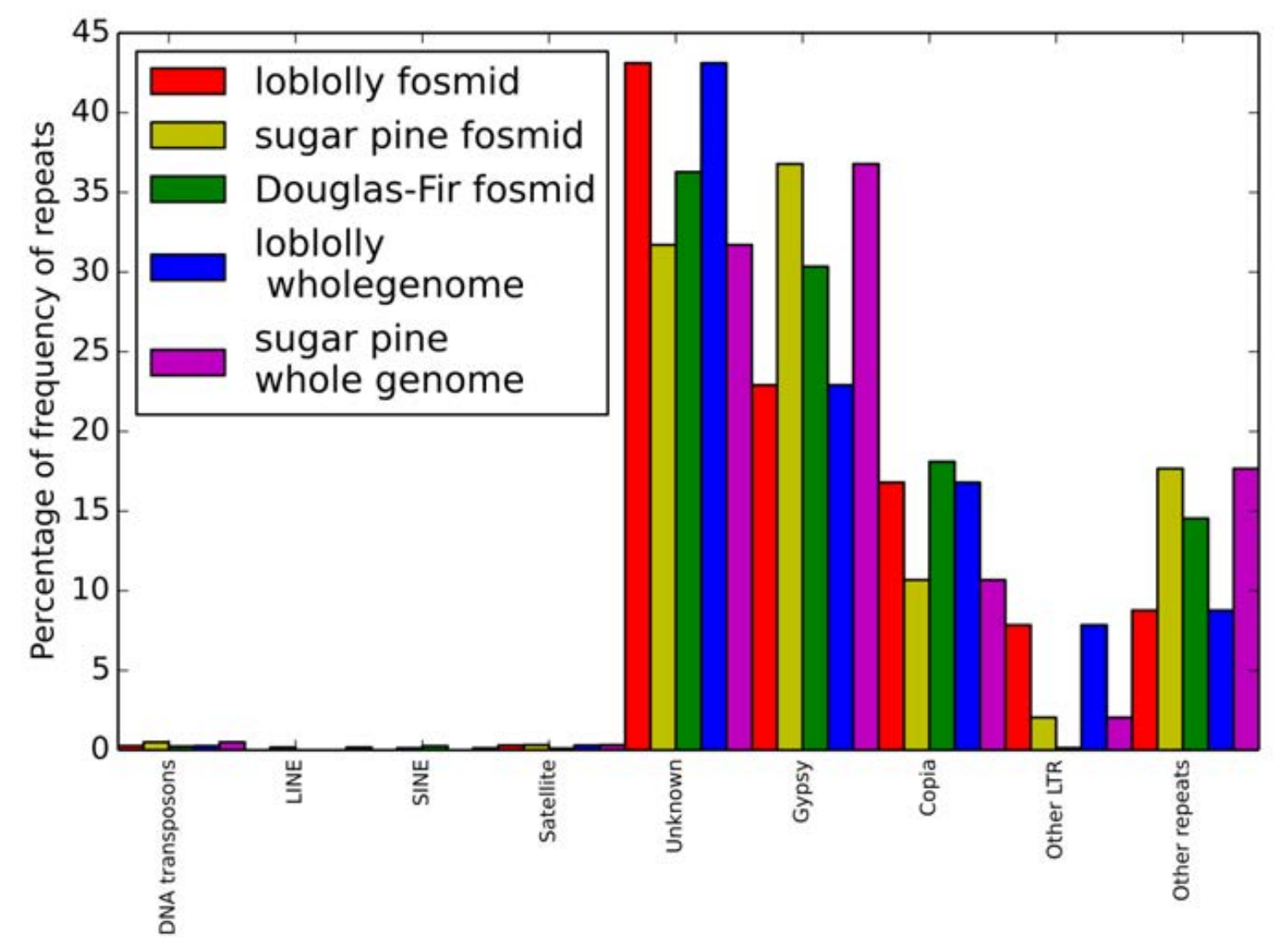

Figure 5: Relative percentage of interspersed repeat categories in various conifer genomic sequence sets

CONCLUSIONS

- **Gymnosperms have not been characterized extensively in existing databases. A combination of *de novo* and library-based approaches are necessary to characterize repeat content in conifers (Figure 1).**
- ***De novo* approaches identified the majority of repeat families however over 30% of these repeats are unclassified and require further characterization (Figure 5).**
- **Both loblolly pine and sugar pine have repetitive content estimates that exceed 75% (Table 2).**
- **Comparative analysis reveals that sugar pine has fewer repeat families that are contributing to large percentage of the interspersed repetitive content than lobloly pine (Figure 4).**

References

- Wegrzyn et al. Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics (2014) 196, 891-909.
- Michael et al. The First 50 Plant Genomes. The plant genome (2013) , 6 (2), 1-7.
- Wegrzyn et al. Insights into the Loblolly Pine Genome: Characterization of BAC and fosmid sequences. Plos One 8 (9), e72439.