Insights into the Loblolly Pine Genome: Characterization of Fosmid Sequences

Jill Wegrzyn
Brian Lin, Jacob Zieve, Matt Dougherty, Pedro Martinez-Garcia, David Neale, Kristian Stevens

Department of Plant Sciences
University of California at Davis

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Characterization of Sequence

- Significant portion of the genome is repetitive
  - Tandem repeats
  - Dispersed elements (transposons and retrotransposons)
- Methodology
  - Similarity search
  - *De novo*
- Gene Content

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Prior Studies in Gymnosperms

Sequence Type	Species	Paper	Repetitive Content	Elements Identified
BACs	Picea glauca	Hamberger et al (2009)	40%	
BACs	Pinus taeda	Morse et al (2009)		Gymny
BACs	Pinus taeda	Kovach et al (2010)	23-80%	PtIFG7
BACs	Taxodium disitchum	Liu et al (2011)	90%	
Fosmids	Taxus mairei	Hao et al (2010)	20.8%	
Southern Hybridization	Pinus pinaster	Rocheta et al (2006)		PpRT1
Southern Hybridization	Pinus elliottii	Kamm et al (1996)		TPE1

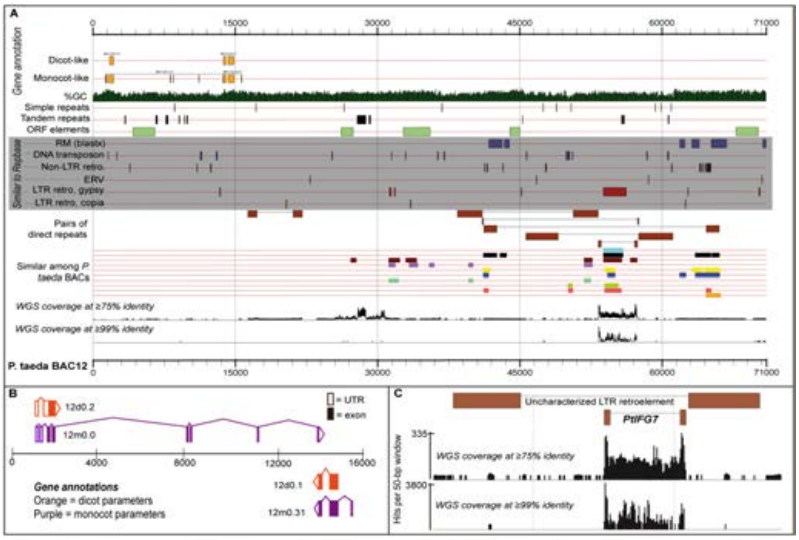
Sequence Data Sets

Pinus taeda BACs and Fosmids

	Pinus taeda BACs	*Pinus taeda* Fosmids
Total number of sequences	103	90,973
Average sequence length	115,130	2,918
Median sequence length	118,782	475
N50 sequence length (bp)	127,167	16,204
Shortest sequence length	1,392	201
Longest sequence length	235,088	75,791
Total length (bp)	11,858,447	265,511,345
GC %	37.98%	38.09%
A : C : T : G%	31.27 : 18.79 : 31.32 : 18.62	30.94:19.07:30.97:19.03

Combined sequence resource represents roughly 1% of the estimated 22 GB genome

Annotated BAC Sequences



USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Similarity Search Analysis

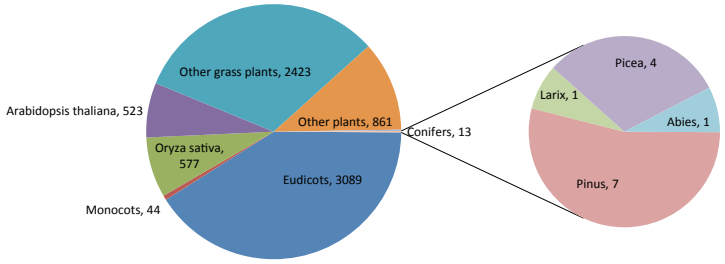
RepBase

- Similarity search approach relies on comparisons against annotated repetitive elements
  - RepBase v. 17.07
    - Number of entries: 29,925
    - Average length of an entry: 2,640 bp
    - Number of species represented: 736
    - Number of repeat families: 320
    - Angiosperm entries: 7,469
    - Gymnosperm entries (conifer): 14 (13)

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Similarity Search Analysis

RepBase



Pinus elliotti	Pinus thunbergii	Pinus radiata	Picea abies	Picea glauca	Abies veitchii	Larix
TY1_PE (Copia-LTR)	PLN1_PT (LINE) PLN2_PT (LINE) PLN3_PT (LINE) RT_PT (Copia-LTR)	IFG7_I (Gypsy-LTR) IFG7_LTR (Gypsy-LTR)	AlISEI_LTR(Gypsy-LTR) ALISEI_I(Gypsy-LTR)	PGGYPSYX1 (Gypsy-LTR) PCOPIAX1 (Copia-LTR)	ROMANIAV1 (Gypsy-LTR)	SAT1_LM (Satellite)

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Custom Sequence Included in RepBase

Class	Order	Superfamily	Name	Accession	Species	Length
I	LTR	Gypsy	PtIFG7	-	Pinus taeda	3660
I	LTR	Gypsy	TPE1	Z50750	Pinus elliottii	1663
I	LTR	Gypsy	IFG7-PpRT1	DQ394069	Pinus pinaster	5966
I	LTR	Gypsy	RLG_GYMNY-1	EU912388.1	Pinus taeda	6113
I	LTR	Gypsy	Corky	EU862277.1	Quercus suber	5924

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

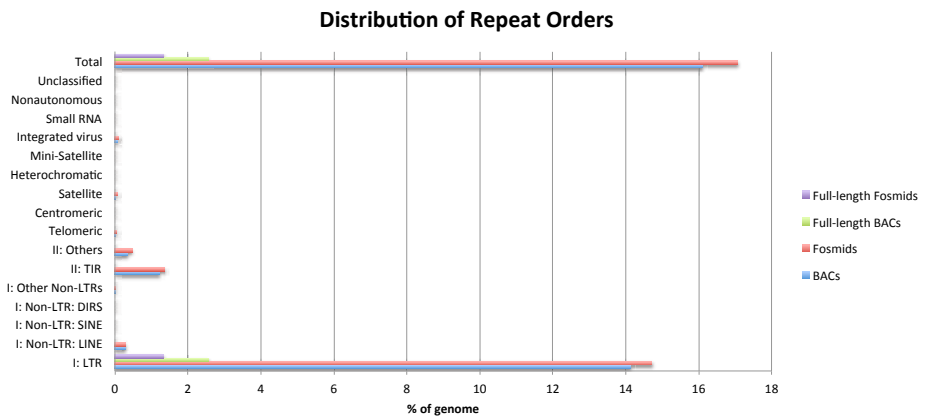
Censor: Library Based Approach

- TE identification and tandem repeats
  - Compare input data to set of reference sequences
  - Can only detect what already exists
  - Modified NCBI BLAST algorithm (local alignment)
  - Homologous bases "censored"
  - CENSOR and BLAST not directly comparable
    - Rough, normal, and sensitive (modes of operation)
  - Different parameters than RepeatMasker
    - More flexibility

Overview of Censor Results

	BACs	Fosmids
Total Alignments	6,167	175,004
Average repeat length(bp):	318.95	267
Average similarity(%):	78.22	78.55
% of genome prior to filtering	16.59%	17.63%
Alignments that pass 80-80-80 rule:	84	989
Average repeat length(bp):	3622.38	3583.6
Average similarity (%):	89.9	88.48
% of genome	2.57%	1.33%
% repeats annotated as full-length	15.47%	7.57%

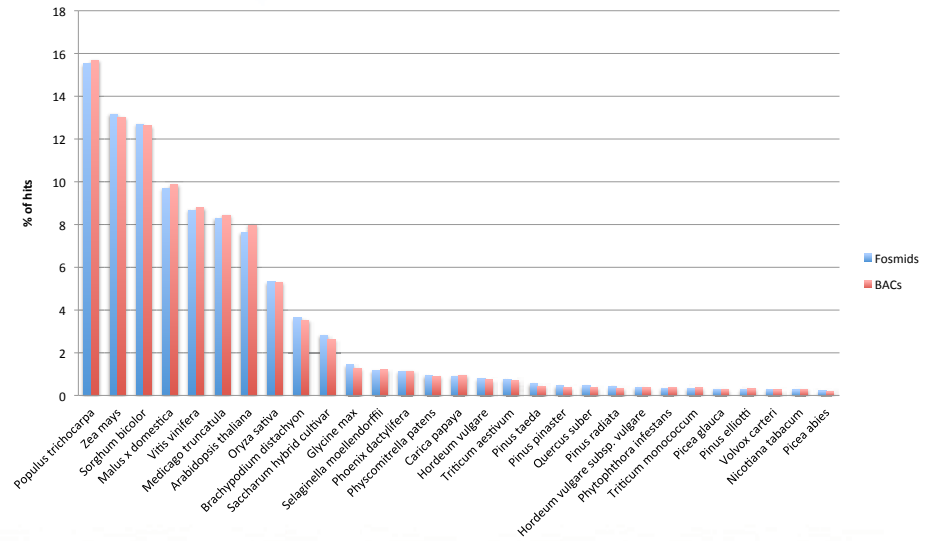
Censor Similarity Search (RepBase)



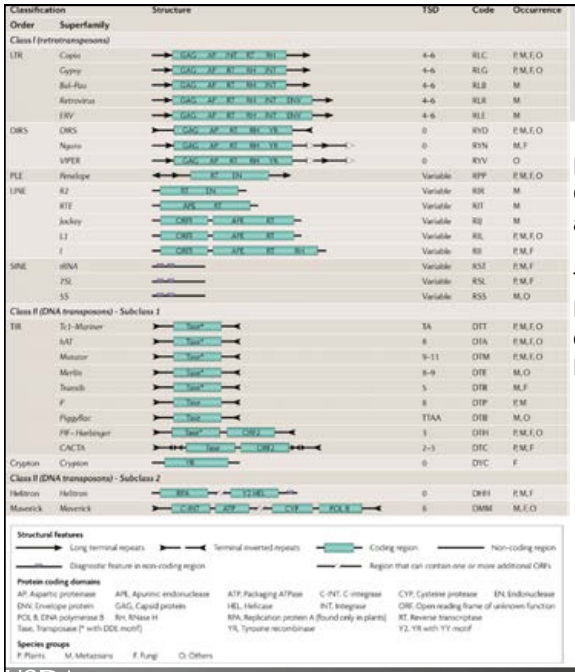
Unfiltered results from Censor against Repbase annotate approximately 20% of the genome versus 2% after the application of the 80-80-80 rule (Wicker et al 2007)

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Distribution by Species (Censor)



USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq



Unified Classification System for Eukaryotic Transposable Elements (Wicker et al. 2007)

Hierarchial classification system: Class, sub-class, order, super-family, and family

Terminal repeat regions and other Non-coding portions are the fastest evolving (most specificity in defining Families)

Super-families of full-length fragments

BACs						
Class	Order	Superfamily	Copy number	Length (bp)	% of repeats	% of genome
I	LTR	Gypsy	46	183907	61.43	1.55
I	LTR	Copia	35	102040	33.53	0.86
I	LTR	Gymny	3	18333	6.03	0.15
Fosmids						
Class	Order	Superfamily	Copy number	Length (bp)	% of repeats	% of genome
I	LTR	Gypsy	586	2188186	61.74	0.82
I	LTR	Copia	361	1117069	31.52	0.42
I	LTR	Gymny	40	238103	6.7	0.09
I	LINE	L1	2	825	0.00023	0.00

Most Prevalent Transposable Elements (TEs) in BACs

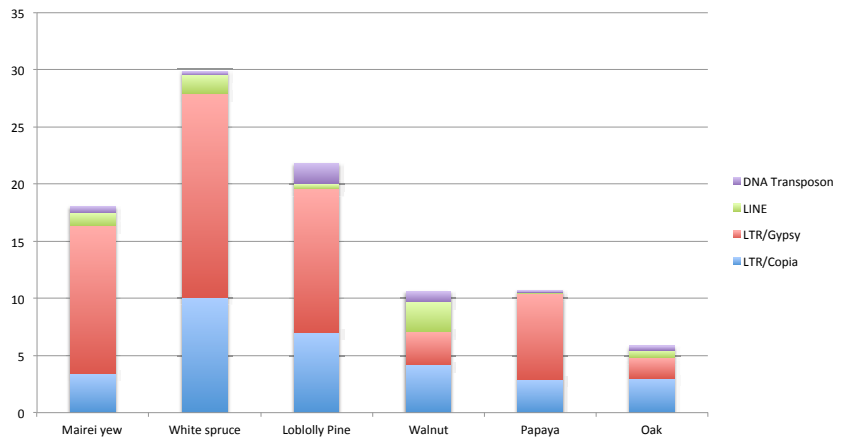
Class	Order	Superfamily	Name	Length (bp)	Copy number	% of repeats	% of genome	Species
I	LTR	Copia	TPE1	99872	26	32.8224	0.84%	Pinus elliottii
I	LTR	Gypsy	Corky	115216	20	37.8651	0.97%	Quercus suber
I	LTR	Gypsy	PGGYPSYX1	44138	20	14.5057	0.37%	Picea glauca
I	LTR	Copia	RT_PT	1926	8	0.6329	0.02%	Pinus thunbergii
I	LTR	Gypsy	PtIFG7	13765	4	4.5237	0.12%	Pinus taeda
I	LTR	Gypsy	RLG_Gymny-1	18333	3	6.025	0.15%	Pinus taeda
I	LTR	Gypsy	IFG7-PpRT1	10788	2	3.545	0.09%	Pinus pinaster
I	LTR	Copia	RT_GB	242	1	0.0795	0.00%	Ginkgo biloba

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

Most Prevalent Transposable Elements (TEs) in Fosmids

Class	Order	Superfamily	Name	Length (bp)	Copy number	% of genome	Species
I	LTR	Copia	TPE1	1080955	287	0.41%	Pinus elliottii
I	LTR	Gypsy	PGGYPSYX1	573079	262	0.22%	Picea glauca
I	LTR	Gypsy	Corky	1033897	178	0.39%	Quercus suber
I	LTR	Gypsy	PtIFG7	259718	77	0.10%	Pinus taeda
I	LTR	Copia	RT_PT	15508	65	0.01%	Pinus thunbergii
I	LTR	Gypsy	RLG_Gymny-1	238103	40	0.09%	Pinus taeda
I	LTR	Gypsy	IFG7-PpRT1	207570	37	0.08%	Pinus pinaster
I	LTR	Gypsy	IFG7_I	110979	23	0.04%	Pinus radiata
I	LTR	Gypsy	IFG7_LTR	2943	9	0.00%	Pinus radiata
I	LTR	Copia	Copia4-PTR_I	19455	4	0.01%	Populus trichocarpa
I	LTR	Copia	RT_GB	697	3	0.00%	Ginkgo biloba
I	LTR	Copia	COPIA_ES	454	2	0.00%	Equisetum scirpoides
I	LINE	L1	PILN1_PT	825	2	0.00%	Pinus thunbergii

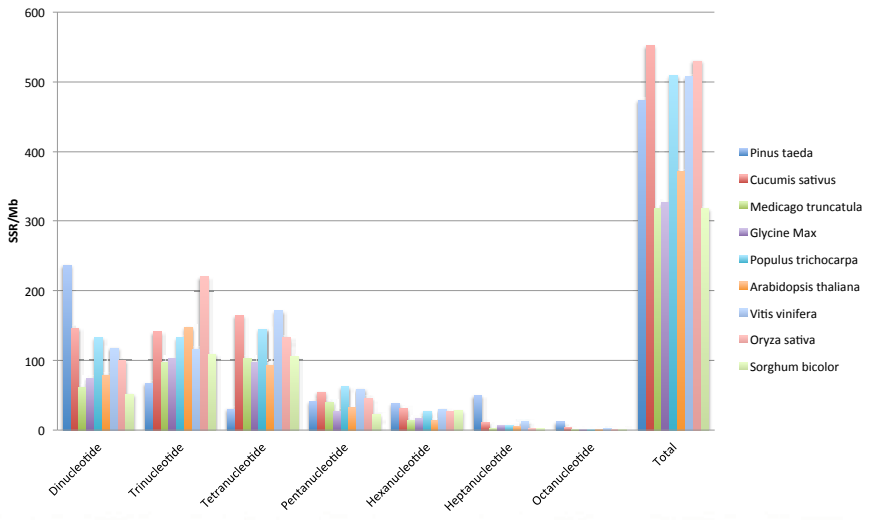
Comparison of TEs across species



Tandem Repeat Analysis (TRF)

	Microsatellite (1-8bp)	Minisatellite (9-100bp)	Satellite (100+bp)	Total
Most frequent period	2	21	123	-
Cumulative bp of the most frequent period	118938	249169	154835	-
Approx. count of the most frequent period	59469	11865	1258	-
Most frequent's % of the genome	0.04	0.09	0.06	-
Most frequent motif	AT	CATATGTCTAAAAATAGA	AAGAAATGAATCCGACCATATCTTGTAAAGCATTGGAAGAAATTACCACTCCTCAAACTCTCAAGATAGAAGGACATATCCAAGAAGAAAAAGGTTCAAGAAGTGAAAGATCACATCGAGCATC	-
Occurrences of motif	946	105	5	-
Total length	358200	4738502	2259658	28056360
% of genome	0.13	1.78	0.85	2.76

Microsatellite/SSR Frequency



USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

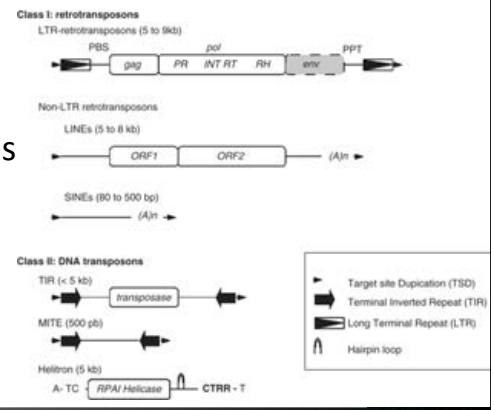
Tandem Repeats (Conifers)

	Pinus taeda BACs			Pinus taeda Fosmids			Taxus mairei Fosmids			Picea glauca BACs		
	Micro	**Mini**	**Sat**	**Micro**	**Mini**	**Sat**	**Micro**	**Mini**	**Sat**	**Micro**	**Mini**	**Sat**
Approximate number	643	3700	412	6213	70762	7339	134	201	16	173	18	8
% of genome	0.34	4.4		0.14	4.16		0.28	1.78		0.81	3.93	
Average length (bp)	63.75	91.75	444.89	61.82	91.16	425.93		103.2			96.5	422.4
Most common period size	**2**	21	123	**2**	21	123	**2**	22	230	**2**	27	113
Longest repeat unit	500			499			359			375		

USDA United States Department of Agriculture National Institute of Food and Agriculture PineRefSeq

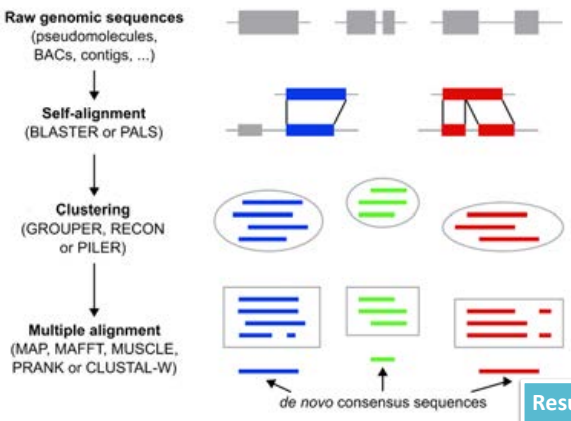
De novo Repeat Identification

- Library-Based (library of features)
  - LTR_STRUCT
- Signature Based
  - Non-LTR Retrotransposans
  - SINDR, FINDMITE
- Self-comparison approaches
  - RECON, Piler, BLASTER
- *Ab initio* (k-mer and seed approaches)
  - RepeatScout, P-Clouds



(Lerat et al. 2010)

REPET Methodology (Tedenovo)

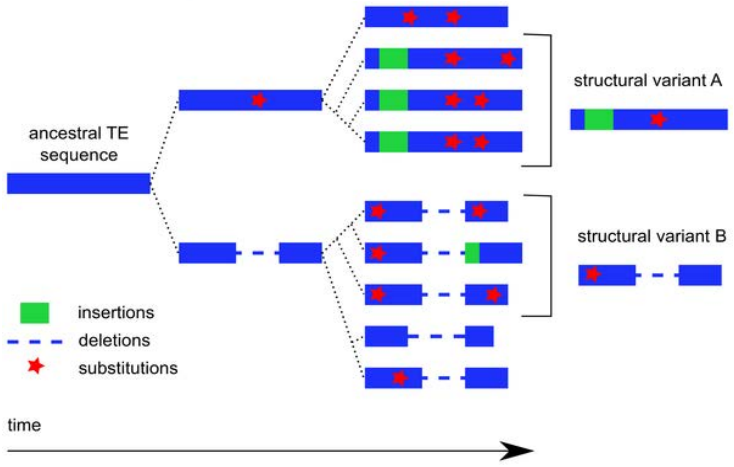


- Self-alignment (all vs all) with BLAST to find HSPs is followed by 3 separate rounds of clustering with Grouper, Recon, and Piler
- 3 sets of clusters are aligned with a MSA, Map, to derive a consensus sequence for each
- Structural search runs simultaneously (LTR Harvest) to detect highly diverged LTRs
- Blastclust to cluster potential sequences

Results from both pathways are combined, and filtered to reduce redundancy between the consensus sequences derived independently

(Flutre et al, 2011)

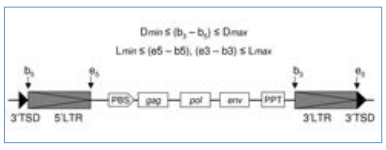
TE Family with two Structural Variants



(Flutre et al, 2011)

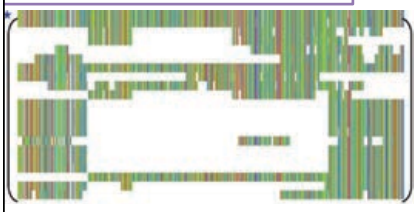
LTRHarvest

- LTR Harvest is designed to detect LTR retrotransposon candidates that contain at least two LTRs
  - Solo LTRs, truncated elements that lack one LTR, or elements with large insertions do not fulfill the model
- Evaluations of applications in this class of structural identification against the Drosophila genome had the best results (Lerat et al. 2010)
- Takes into account several structural features:
  - Size range of the LTR
  - Distance between two LTRs
  - Presence of TSDs at each extremity
  - Replication sites (primer binding site and polypurine tract)
  - Percentage identity between 2 LTRs
  - Presence of conserved motifs corresponding to genes encoded



REPET Methodology (Classification)

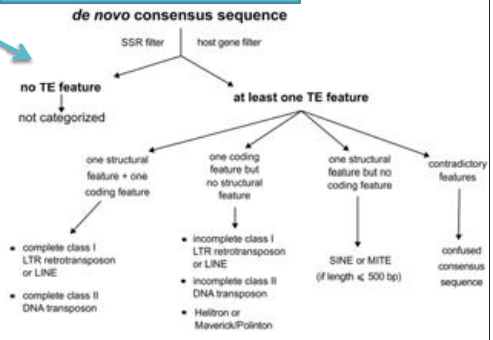
DNA TE (3.5kbp TIR) discovered by RECON clustering and MAP alignment



- Repbase was used as the repeat element library
- HMMER profiles formatted for repeats was used in conjunction with HMMER 3.0
- Host gene database was curated and constructed with 560 full-length loblolly pine cDNAs and 945 Pinus radiata full-length cDNAs
- Ribosomal DNA database was constructed from 52 curated *Pinus* sequences.
- Poly-A tails, tandem repeats, open reading frames, and terminal repeats are also detected
- The repeat classification system described in Wicker et al. (2007) is applied

- Consensus sequences derived from REPET are classified via Teclassifier
- Blasted against several databases for the purpose of order-level assignment

Feature based classification



(Flutre et al, 2011)

PineRefSeq

Full Summary of REPET Results

	BAC	Fosmid
Unique elements	591	11,631
Fragments in dataset	2,444	153,242
Average fragment length	3,364	1,842
combined consensus length (consensus only) (bp)	3,495,285	54,121,238
Repeat length (*de novo*) (bp)	3,905,885	66,188,613
Percentage (*de novo*)	32.83%	23.91%
Repeat length (similarity + *de novo*) (bp)	2,429,433	35,222,731
Percentage (similarity + *de novo*)	62.20%	53.22%
Percentage (*de novo* only)	20.49%	12.72%

USDA United States Department of Agriculture National Institute of Food and Agriculture

PineRefSeq

REPET Classification (by Order)

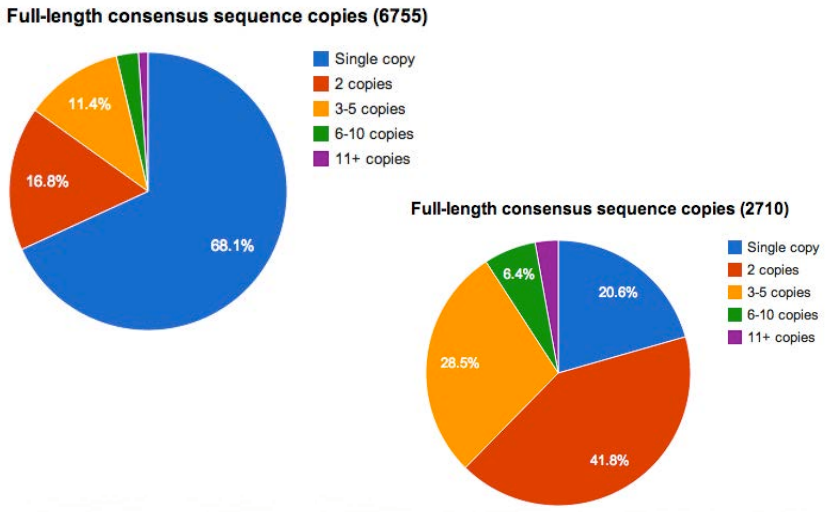
Class	Order/Structure	# consensus sequences (BAC)	% genome (BAC)	# consensus sequences (Fosmid)	% genome (Fosmid)
Class I Retrotransposon	LTR	325	21.57%	5,061	13.38%
	DIRS	2	0.19%	82	0.18%
	LINE	16	0.81%	274	0.43%
	SINE	2	0.01%	20	0.01%
	TIR	11	0.58%	122	0.17%
	Helitron	0	0%	12	0.02%
	LARD	117	5.09%	2,068	3.81%
	TRIM	15	0.22%	244	0.17%
	All Class I	488	28.65%	8,097	18.74%
Class II DNA transposon	MITE	2	0.08%	12	0.00%
	All Class II	13	0.65%	155	0.22%
Uncategorized		81	3.25%	2,953	4.55%
Total		591	32.83%	11,631	23.91%

Comparison between REPET stages

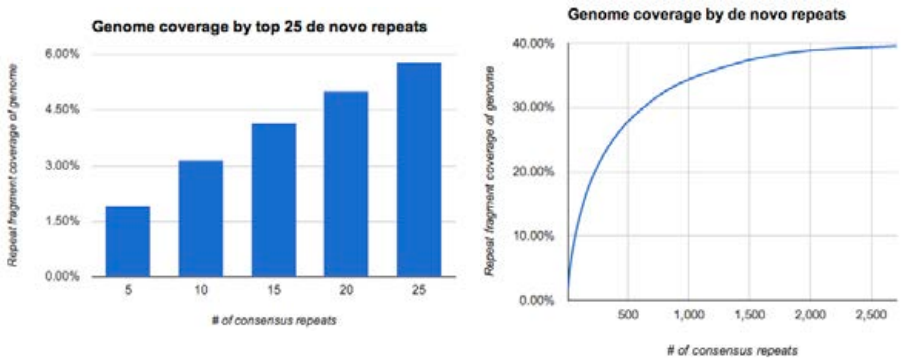
	Total	LTRharvest	Grouper, Recon, Piler
Consensus sequences (UClust of 12,000)	7757	5386	2371
Full-length sequences (Post-mapping)	6755	4995	**1760**
Single copy full-length sequences	4603	4045	558
2+ copy full-length sequences	2152	**950**	1202
Average sequence length (bp)	4709	6053	2654

The high number of single-copy repeats identified by LTR-harvest led to more stringent criteria for positively identifying repeats in the sequence.

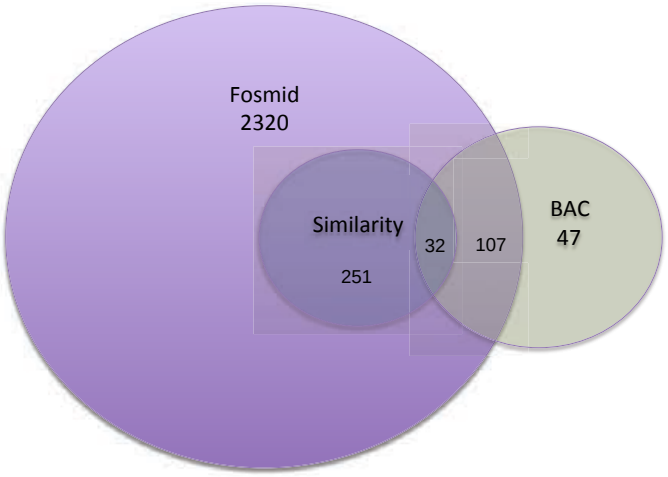
Full-length Consensus Sequences



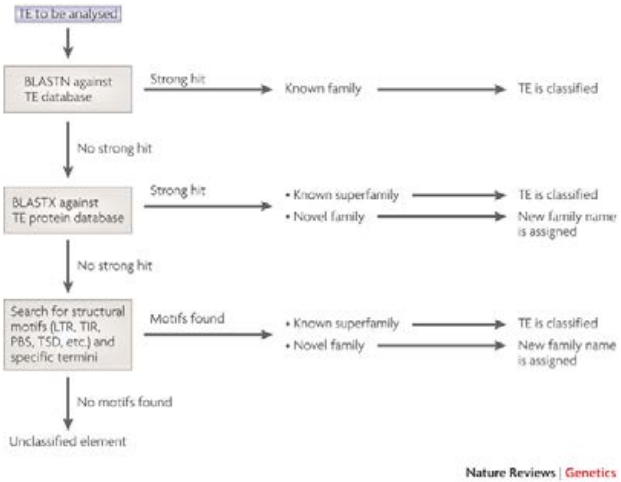
Most Prevalent Repeats



TE Identification



TE Classification



High Copy Number Elements

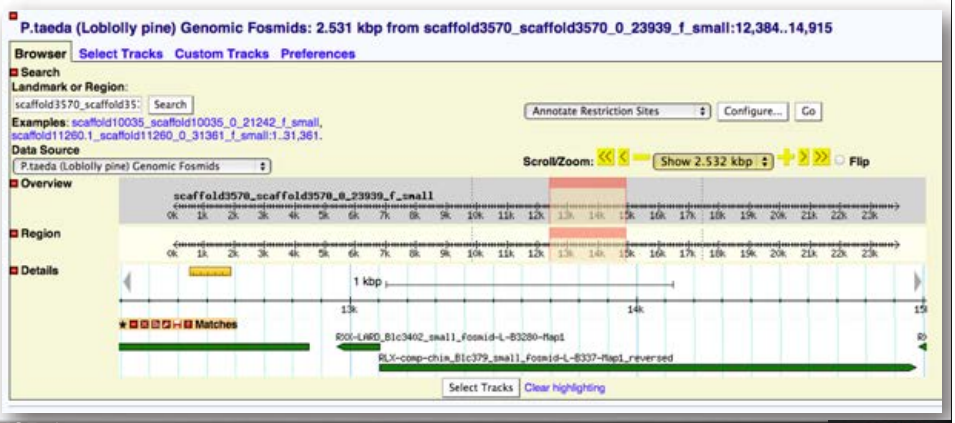
Name	# copies	Total bps	Annotation
G11355	37	427637	Class I : LTR : Gypsy
B2962	45	256184	Class I : LTR : Gypsy: IFG7 : PtIFG7
R3	40	201109	Class I : LTR : Copia
R243	24	184879	Class I : LTR
B3068	33	184476	Class I : LTR : Gypsy: Gymny
B1892	23	170184	Class I
B4275	33	141121	Class I : LTR : Copia
B3021	25	140970	Class I : LTR : Gypsy : IFG7 : IFG7_I
B1613	20	147634	Class I (LARD)
G11339	18	156793	Class I : LTR : Gypsy: Corky
B141	18	251535	Class I : LTR : Gypsy
B894	18	179194	Class I : LTR
B746	17	178696	Class I : LTR : Gypsy: TPE1
B2	22	128333	Class I : LTR
B2977	22	119426	Class I : LTR
R104	38	118642	Class I : LTR

Gene Content

- Methodology:
  - JigSaw (training set: 560 full length genes)
  - Exonerate (produces a spliced alignment)
  - Alignments constructed with curated plant protein and RNA
- Results:
  - Gene prediction estimates 2% gene content
    - Rejects most models as pseudogenes
  - Validated gene prediction estimates of 1.1% (920 genes)

Data Availability

- Alignment of consensus repeats against the fosmid sequences implemented (GFF3)
- Fosmids will be made available in Genbank
- Repeat sequences submitted to Repbase



Acknowledgements

Repeat Team - UCDavis
Department of Plant Sciences
Brian Lin
Jacob Zieve
Pedro Martinez-Garcia
David Neale

Department of Evolution and Ecology
Kristian Stevens
Matt Dougherty