The repetitive DNA of conifers

- a snapshot of the *Piceaabies*(Norway spruce) genome and re-sequencing of 5 other conifers

Anna Wetterbom
Science for Life Laboratory
Karolinska Institute, Sweden

Andrea Zuccolo, Applied Genomics Institute, Italy
Manfred Grabherr, Uppsala university, Sweden
Carlos Talavera-Lopez, Science for Life Laboratory, Sweden

SciLifeLab

The repeat landscape in *Piceaabies* (so far)

- 75 % of genome is repetitive DNA
  - Re-association kinetics (e.g. Rake, 1980)
- Transposable elements are a large part
- Repbase has only 13 sequences from coniferales

The repeat landscape in *Piceaabies* (so far)

Classification	N. of occurrences	Genomic sequence fraction (%)
Repeated sequences	**8822**	**77.44**
Transposable elements	4902	43.03
Class I (retrotransposons)	4834	42.43
Order LINE	151	1.33
Order LTR	4683	41.11
Superfamily Copia	1483	13.02
Superfamily Gypsy	2861	25.11
Superfamily Retrovirus	15	0.13
Unknown superfamily	324	2.84
Class II (DNA transposons)	68	0.60
Subclass 1		
Superfamily CACTA	27	0.24
Superfamily hAT	24	0.21
Unknown superfamily	2	0.02
Subclass 2		
Superfamily Helitron	7	0.06
Tandem repeats	74	0.65
Tandem repeats	7	0.06
rDNA	75	0.66
Unclassified repeats	3846	33.76

Emanuele De Paoli et al. In preparation

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content
- …

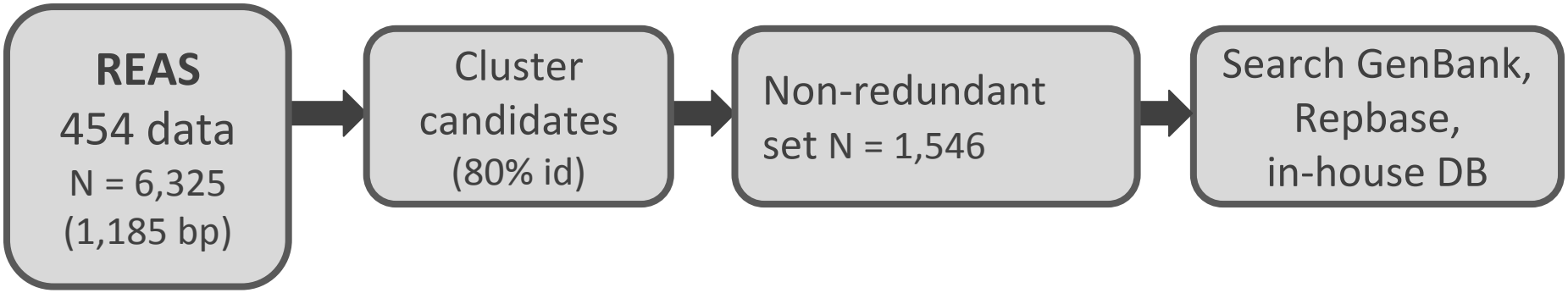
Why study Transposable Elements (TEs)?

- Constitute a significant portion of many eukaryotic genomes (e.g. >85% in maize)
- Mutagenic
  - Inactivate genes
  - Affect transcription
- Grab, duplicate and transfer part of genes
  - Helitrons, Pack MULE
- Used in phylogenetic studies
- Valuable biotechnological tool
  - Use to tag genes

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization **How?**
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content

Identification of repeats, mainly TEs



Identification of repeats, mainly TEs

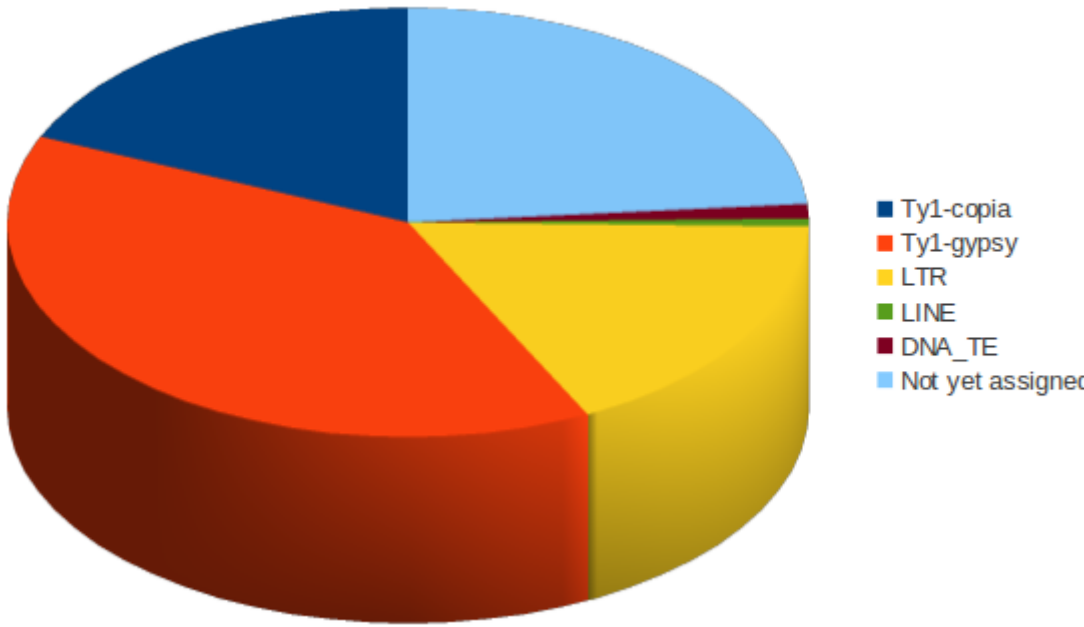
REAS
454 data
N = 6,325
(1,185 bp)

→ Cluster candidates (80% id)

→ Non-redundant set N = 1,546

→ Search GenBank, Repbase, in-house DB

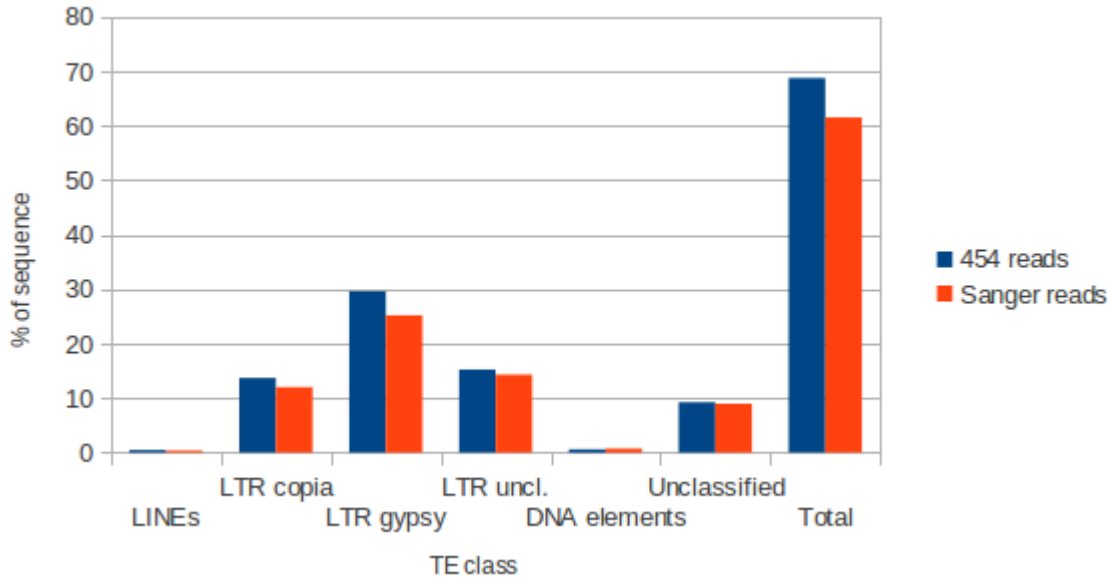
```
DNA_TE:              19
LINE:                13
LTR:
    unknown:         265
    Ty1-copia:       265
    Ty3-gypsy:       591
Unclassified:        393
```



Quantification

Using Repeat Masker and custom repeat library.

	454 reads	Sanger RS
LINE	0.44	0.35
LTR copia	13.69	11.99
LTR gypsy	29.61	25.19
LTR uncl.	15.22	14.3
DNA_TEs	0.57	0.72
Unclassified	9.24	8.97
Total	68.77	61.52



What about repeats in the genome assembly?

Diploid WGS assembly

- 50x genome coverage
  - Illumina PE, 100 bp reads
  - 180 bp, 300bp and 650 bp libraries
- 1.7x genome coverage
  - 454 SE, 450 and 700 bp reads

Assemble with CLC

Assembly stats:

- 44 % of genome in contigs> 1 kbp
- 12 % in contigs> 5 kbp
- 3 % in contigs> 10 kbp
- NG50: 757 bp

What about repeats in the genome assembly?

Diploid WGS assembly

- 50x genome coverage
  - Illumina PE, 100 bp reads
  - 180 bp, 300bp and 650 bp libraries
- 1.7x genome coverage
  - 454 SE, 450 and 700 bp reads

Assemble with CLC

Assembly stats:

- 44 % of genome in contigs> 1 kbp
- 12 % in contigs> 5 kbp
- 3 % in contigs> 10 kbp
- NG50: 757 bp

Repeat content in the diploid WGS assembly

	Contigs 0.2 - 1 kbp (N=1,281,983)	Contigs 1-3 kbp (N=683,251)	Contigs 3-5 kbp (N=318,889)	Contigs > 5 kbp (N=500)
SINEs	0	0	0	0
LINEs	2.2 %	0.7 %	0.7 %	0.6 %
LTR elements	70 %	50 %	47 %	11 %
DNA elements	1.2 %	0.9 %	0.9 %	0.3 %
Unclassified	12 %	8.8 %	8.0 %	2.3 %
Total interspersed repeats	85 %	60 %	75 %	14 %

Repeat Masker with custom libraries

Repeat content in the diploid WGS assembly

	Contigs 0.2 - 1 kbp (N=1,281,983)	Contigs 1-3 kbp (N=683,251)	Contigs 3-5 kbp (N=318,889)	Contigs > 5 kbp (N=500)
SINEs	0	0	0	0
LINEs	2.2 %	0.7 %	0.7 %	0.6 %
LTR elements	70 %	50 %	47 %	11 %
DNA elements	1.2 %	0.9 %	0.9 %	0.3 %
Unclassified	12 %	8.8 %	8.0 %	2.3 %
Total interspersed repeats	85 %	60 %	75 %	14 %
Simple repeats*	0.8 %	0.2 %	0.2 %	0.2 %
Low complexity regions*	2.2 %	1.0 %	01.0 %	1.5 %
Ribosomal repeats**	2.1 %	1.1 %	1.2 %	1.2 %

Repeat Masker with custom libraries

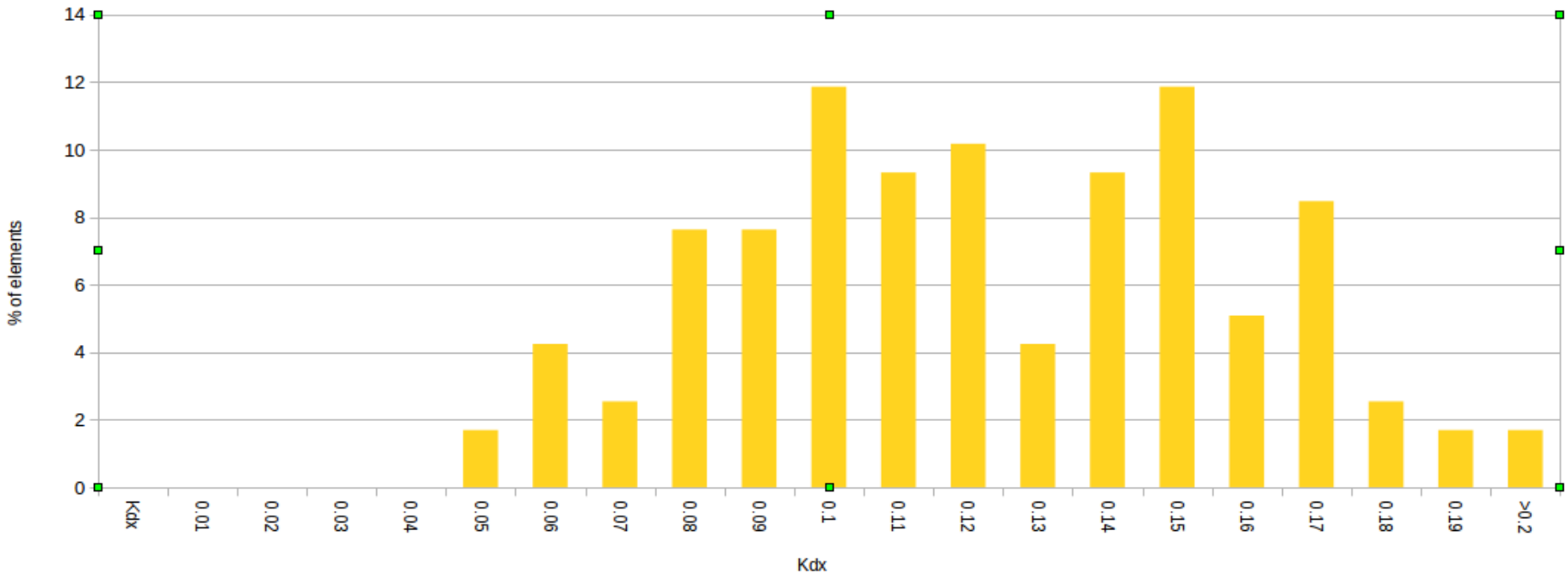
* From Repbase (plant section)

** From GenBank (*P. abies*ribo.seq)

Roadmap: the repetitive fraction of the Norway spruce genome

- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content

Insertion time of LTR-RTs



4.24 Million years 84.75

For each of the complete 126 LTR-RTs identified we calculated the nucleotide distance between their LTRs using the Kimura 2 param. method.
Assuming a mutation rate of 2.36×10^{-9} most of the elements appear to be quite old.

What about uncharacterized candidates?

What could they be?

LTR regions, MITEs and non autonomous DNA elements, Helitrons, satellite repeats...

How can they be identified?

Using a mix of searches for structural features and motif searches

Roadmap: the repetitive fraction of the Norway spruce genome

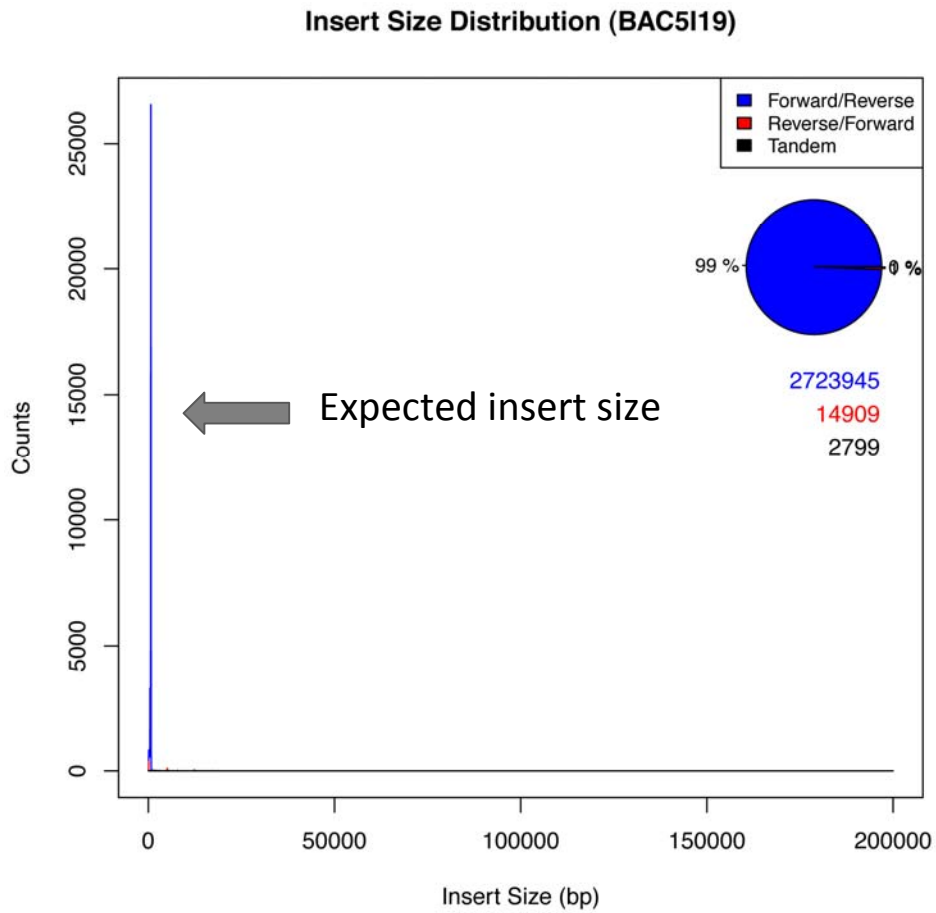
- Classify repeats and create a comprehensive repeat library for annotation of the assembly
  - Identification
  - Characterization
  - Quantification
- Examine age of different classes of elements
  - Comparative studies
- Examine the importance of TEs for the large genome size
  - Comparative studies
- Examine within species variation in repeat content
- …

Intraspecies variation in *P.abies*

- 4*P. abies*BACs in GenBank
- Map our spruce PE data
  - 650 bp library

Intraspecies variation in *P.abies*

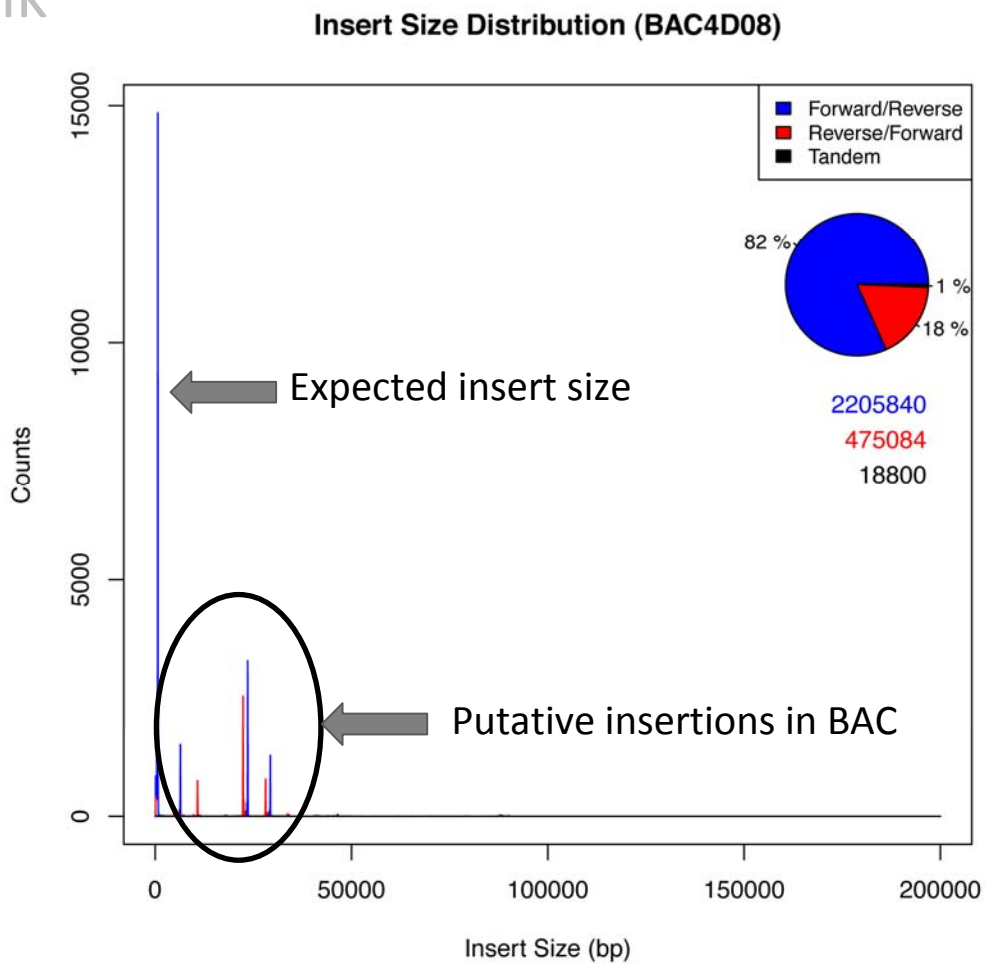
- 4*P. abies*BACs in GenBank
- Map our spruce PE data
  - 650 bp library



Intraspecies variation in *P.abies*

- 4*P. abies*BACs in GenBank
- Map our spruce PE data
  - 650 bp library

- Further investigation of BACs
- Scale up: re-sequence five more individuals
  - Low coverage

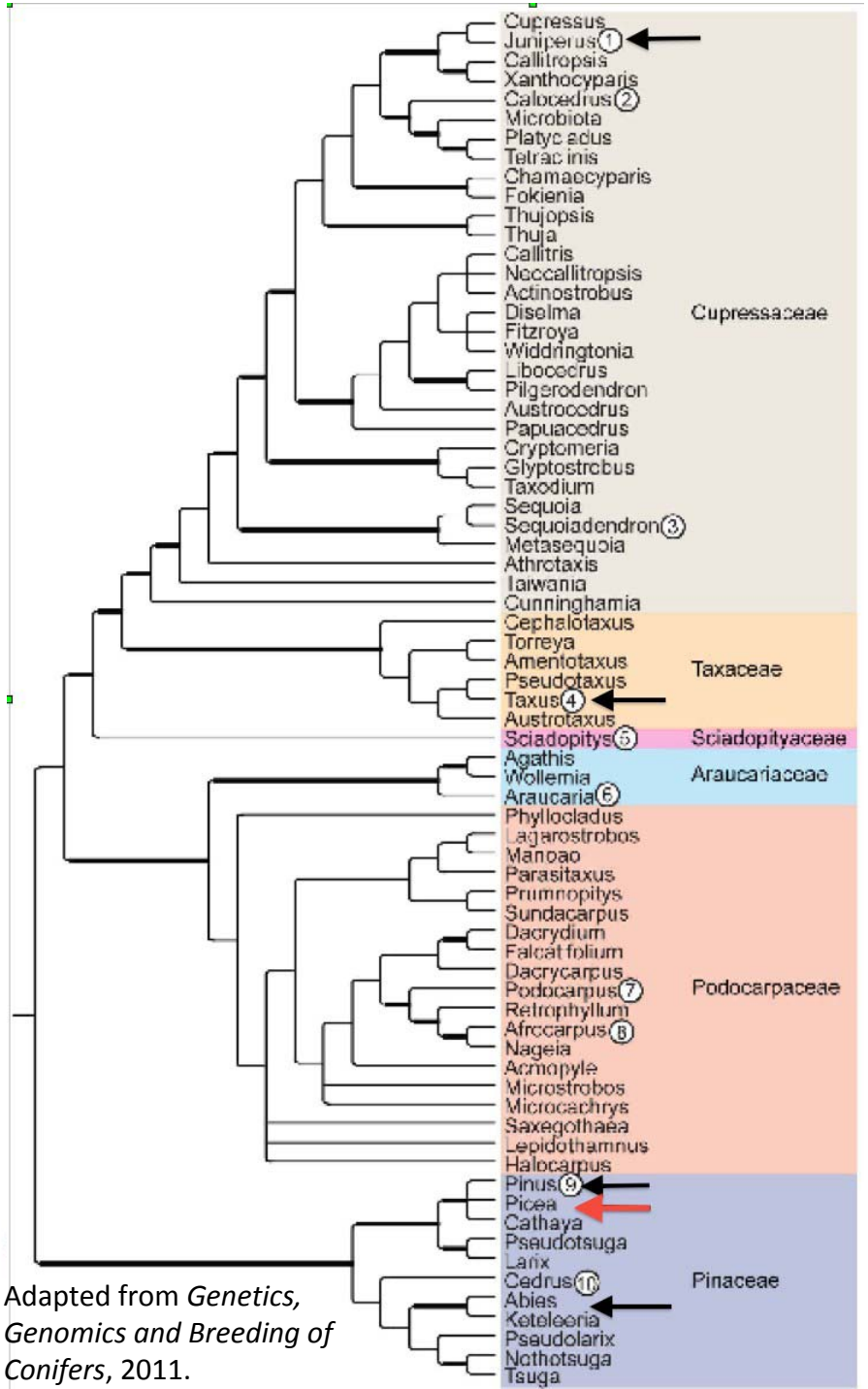


Comparative sequencing

When one genome is not enough…

Why sequence more species?

- Time TE colonization of conifer genomes
- Estimate TE mutation rate
- Investigate the high degree of nucleotide sequence conservation across genera for certain TEs
- Search for species specific TEs and repetitive sequences



Adapted from *Genetics, Genomics and Breeding of Conifers*, 2011.

Piceaabies

Pinussylvestris

Abiessibirica

Taxusbaccata

Junpieruscommunis

Gnetum

Sequencing strategy

Illumina PE

- 20x genome coverage for pine ✓
- 5x genome coverage for others ✓

454 SE

- 100,000 reads (700 bps)

Approaches for analyzing repeats in other conifers

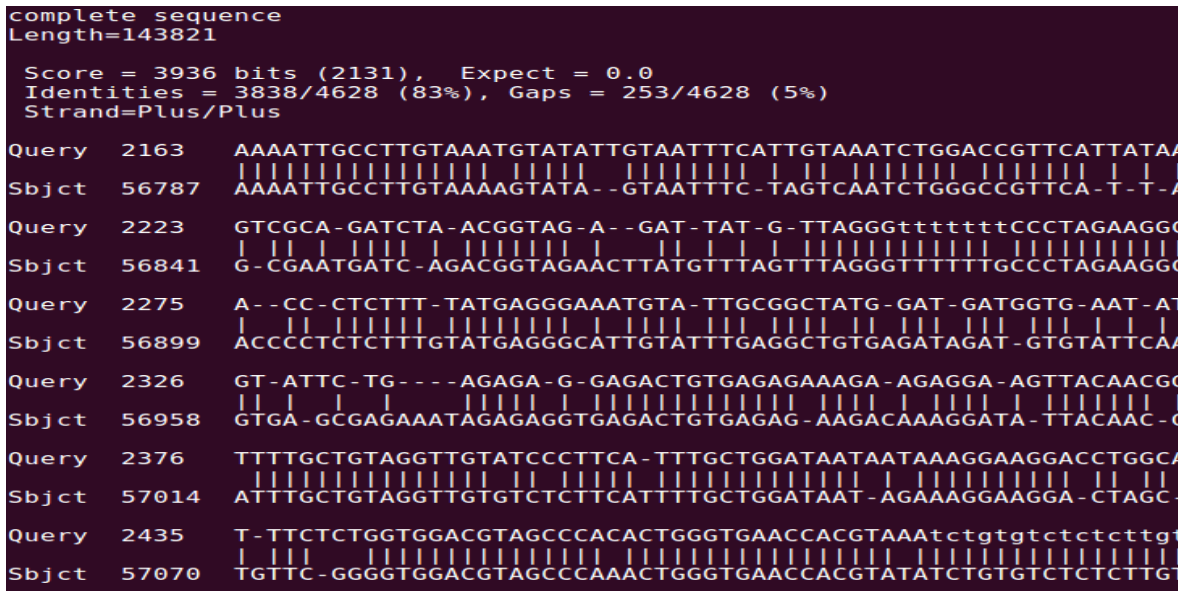
- Mapping reads to the spruce repeat library
  - Problems: mapping parameters vs sequence divergence
- Using the spruce repeat library to mask other assemblies
- *De novo* assembly of repeats
  - using Inchworm
- *De novo* repeat libraries using 454 data
  - REAS and manual curation (similar to spruce)

Approaches for analyzing repeats in other conifers

- Mapping reads to the spruce repeat library
  - Problems: mapping parameters vs sequence divergence
- Using the spruce repeat library to mask other assemblies
- *De novo* assembly of repeats
  - using Inchworm
- *De novo* repeat libraries using 454 data
  - REAS and manual curation (similar to spruce)

Strong sequence conservation across genera

- 111 *Pinustaeda*BACs (12 Mbp)
- TEsidentified in *P. abies* mask 28 % of BACs
- Average similarity is 76 %



```
complete sequence
Length=143821

 Score = 3936 bits (2131),  Expect = 0.0
 Identities = 3838/4628 (83%), Gaps = 253/4628 (5%)
 Strand=Plus/Plus

Query  2163   AAAATTGCCTTGTAAATGTATATTGTAATTTCATTGTAAATCTGGACCGTTCATTATAA
              ||||||||||||||||| |||||  |||||||| | || |||||||| |||||||| | | |
Sbjct  56787  AAAATTGCCTTGTAAAAGTATA--GTAATTTC-TAGTCAATCTGGGCCGTTCA-T-T-A

Query  2223   GTCGCA-GATCTA-ACGGTAG-A--GAT-TAT-G-TTAGGGttttttttCCCTAGAAGGC
              | || | |||| | |||||||| |   || | | | |||||||||||||| ||||||||||
Sbjct  56841  G-CGAATGATC-AGACGGTAGAACTTATGTTTAGTTTAGGGTTTTTTGCCCTAGAAGGC

Query  2275   A--CC-CTCTTT-TATGAGGGAAATGTA-TTGCGGCTATG-GAT-GATGGTG-AAT-AT
              |  || |||||| |||||||| | |||| ||| |||| || ||| ||| ||| | | | |
Sbjct  56899  ACCCCTCTCTTTGTATGAGGGCATTGTATTTGAGGCTGTGAGATAGAT-GTGTATTCAA

Query  2326   GT-ATTC-TG----AGAGA-G-GAGACTGTGAGAGAAAGA-AGAGGA-AGTTACAACGC
              || |  | |     |||||| | |||||||||||||| |||| | |||| | |||||||
Sbjct  56958  GTGA-GCGAGAAATAGAGAGGTGAGACTGTGAGAG-AAGACAAAGGATA-TTACAAC-G

Query  2376   TTTTGCTGTAGGTTGTATCCCTTCA-TTTGCTGGATAATAATAAAGGAAGGACCTGGCA
              ||||||||||||||| || ||||| |||||||||||||| | |||||||||| || ||
Sbjct  57014  ATTTGCTGTAGGTTGTGTCTCTTCATTTTGCTGGATAAT-AGAAAGGAAGGA-CTAGC-

Query  2435   T-TTCTCTGGTGGACGTAGCCCACACTGGGTGAACCACGTAAAtctgtgtctctcttgt
              | |||   |||||||||||||||| |||||||||||||||| ||||||||||||||||
Sbjct  57070  TGTTC-GGGGTGGACGTAGCCCAAACTGGGTGAACCACGTATATCTGTGTCTCTCTTGT
```

Approaches for analyzing repeats in other conifers

- Mapping reads to the spruce repeat library
  - Problems: mapping parameters vs sequence divergence
- Using the spruce repeat library to mask other assemblies
- *De novo* assembly of repeats
  - using Inchworm
- *De novo* repeat libraries using 454 data
  - REAS and manual curation (similar to spruce)

Conclusions

- TEsrepresent the majorityof*P.abies*genomesequence
- LTR retroelementsis the mostabundant TE class
  - Mostof LTR-RT activity in *P.abies*seemstobequiteold.
  - No recentinsertionsweredetected.
  - Some *P. abies*LTR-RT elements are quiteancient in theiroriginpredating the speciationevents.
  - Theseelements show a surprisingly high degreeofsequenceconservationover long evolutionarytimes.
- All the other TE classescouldbeidentified in *P.abies*genome.
- Comparative studiesto date events
  - Mutation rate.
  - Sequencedivergence.

Conclusions

- TEsrepresent the majorityof*P.abies*genomesequence
- LTR retroelementsismostabundant TE class
  - Mostof LTR-RT activity in *P.abies*seemstobequiteold.
  - No recentinsertionsdetected.
  - Some *P. abies*LTR-RT elements are quiteancient in theirorigin.
  - Theseelements show a surprisingly high degreeofsequenceconservationover long evolutionarytimes.
- All the other TE classescouldbeidentified in *P.abies*genome.
- Comparative studiesto date events
  - Mutation rate.
  - Sequencedivergence.

Conclusions

- TEsrepresent the majorityof*P.abies*genomesequence
- LTR retroelementsis the mostabundant TE class
  - Mostof LTR-RT activity in *P.abies*seemstobequiteold.
  - No recentinsertionsweredetected.
  - Some *P. abies*LTR-RT elements are quiteancient in theiroriginpredating the speciationevents.
  - Theseelements show a surprisingly high degreeofsequenceconservationover long evolutionarytimes.
- All the other TE classescouldbeidentified in *P.abies*genome.
- Comparative studiesto date events
  - Mutation rate.
  - Sequencedivergence.

Conclusions

- TEsrepresent the majorityof*P.abies*genomesequence
- LTR retroelementsis the mostabundant TE class
  - Mostof LTR-RT activity in *P.abies*seemstobequiteold.
  - No recentinsertionsweredetected.
  - Some *P. abies*LTR-RT elements are quiteancient in theiroriginpredating the speciationevents.
  - Theseelements show a surprisingly high degreeofsequenceconservationover long evolutionarytimes.
- All the other TE classescouldbeidentified in *P.abies*genome.
- Comparative studiesto date events
  - Mutation rate.
  - Sequencedivergence.

The Spruce Genome Team

UPSC
Rishikesh Bhalerao
Simon Birve
Ulrika Egertsdotter
Ioana Gaboreanu
Rosario Garcia-Gil
Per Gardeström
Thomas Hiltonen
Torgeir Hvidsten
Pär Ingvarsson
Stefan Jansson
Olivier Keech
Susanne Larsson
Chanaka Mannapperuma
Ove Nilsson
Douglas Scofield
Nathaniel Street
Björn Sundberg
Stacey Lee Thompson
Harry Wu

SAB
Kerstin Lindblad-Toh
John MacKay
Outi Savolainen
Detlef Weigel

VIB Gent
Yves Van de Peer
Yao-Cheng Lin

IGA Udine
Michele Morgante
Francesco Vezzi
Ricardo Vicedomini
Andrea Zuccolo

CHORI Oakland
Pieter de Jong
Maxim Koriabine

Skogforsk
Bengt Andersson
Bo Karlsson

SNIC Supercomputers
Uppmax/PDC/NSC/HPC2N

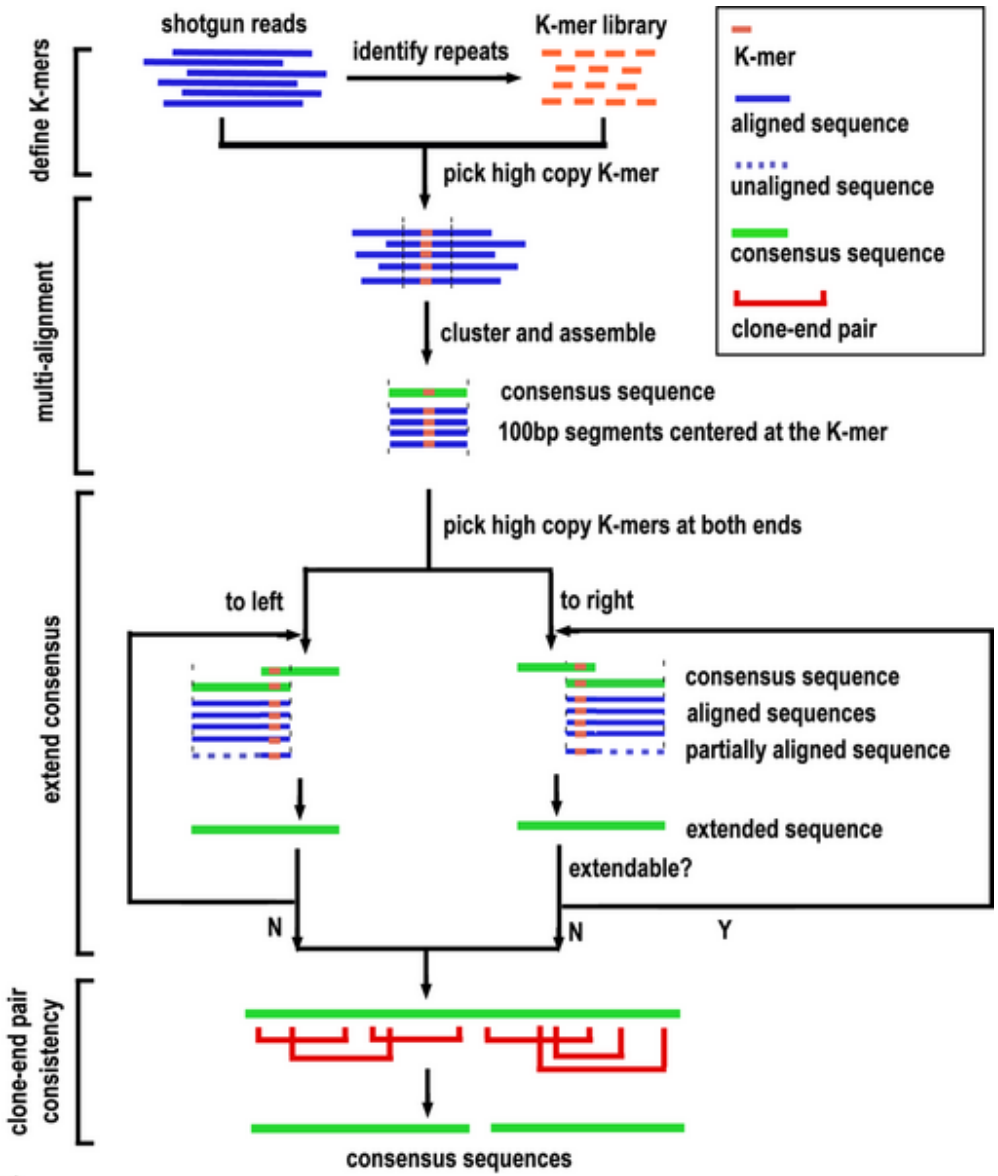
SNISS national infrastructure

SciLifeLab
Andrey Alexeyenko
Björn Andersson
Siv Andersson
Lars Arvestad
Frida Berglund
Oscar Franzén
Manfred Grabherr
Kicki Holmberg
Lisa Klasson
Max Käller
Joakim Lundeberg
Fredrik Lysholm
Björn Nystedt
Kristoffer Sahlin
Ellen Sherwood
Anna Sköllermo
Anne-Charlotte Sonnhammer
Thomas Svensson
Carlos Talavera-Lopez
Anna Wetterbom

CLCbio

Lucigen


The spruce
Thanks for listening!



Ruiqiang Li et al. Sep 2005
PLOS Computational Biology

a) Computing *K*-mer depth: the number of times that a *K*-mer appears in the shotgun data.

b) Seed the process using a randomly chosen high-depth *K*-mer. All shotgun reads containing this *K*-mer are retrieved and trimmed into 100-bp segments centered at that *K*-mer. When the sequence identity between them exceeds a preset threshold, they are assembled into an initial consensus sequence (ICS) using ClustalW.

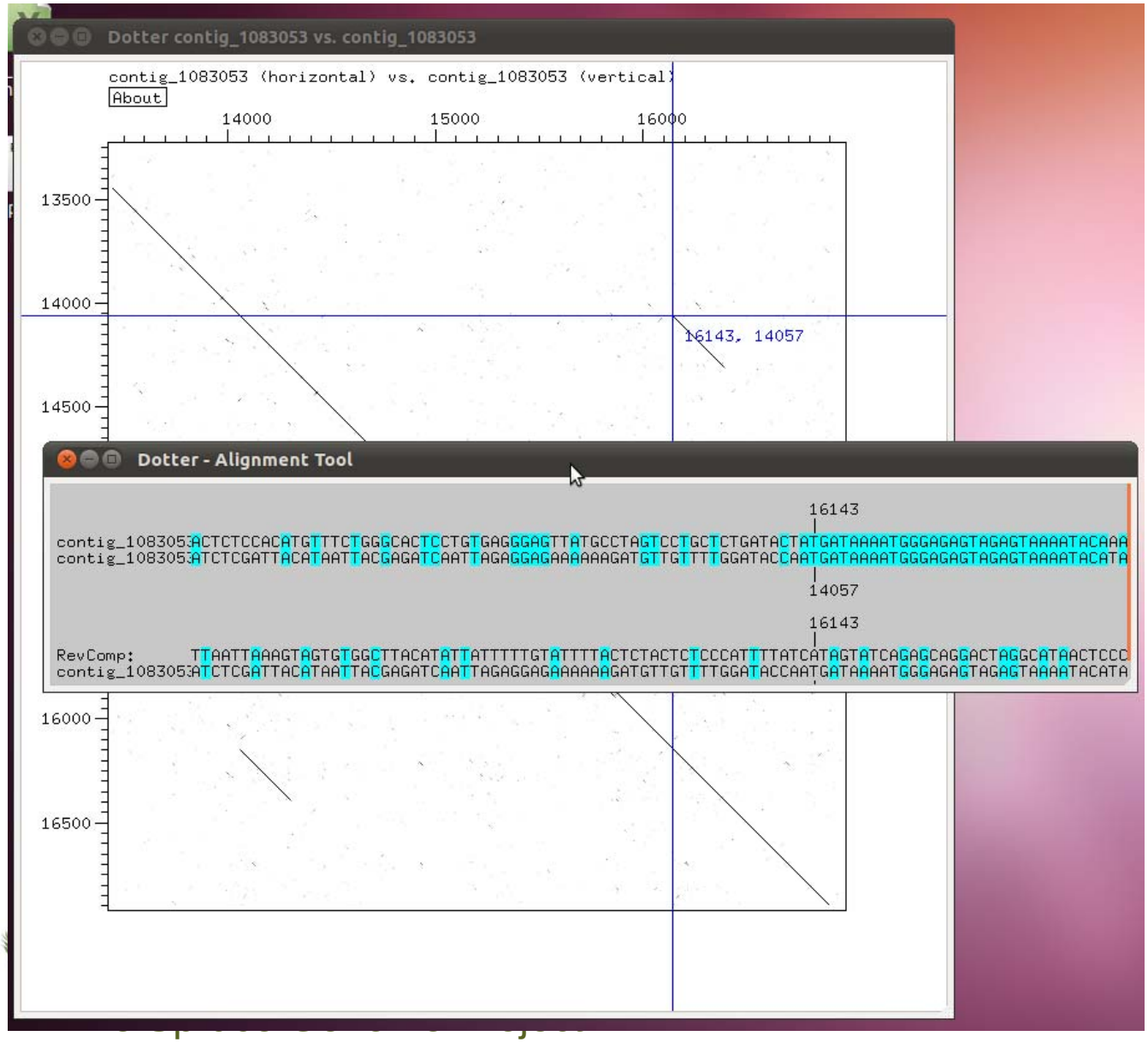
c) An iterative extension is carried out by selecting high-depth *K*-mers at both ends of the ICS and repeating the above procedure.

d) After all such extensions are done, clone-end pairing information is used to resolve ambiguous joins and to break misassembles, but not to join fragmented assemblies. The final consensus is our ReAS (Recovery of Ancestral Sequences) TE.

Identification: a limited pilot test...

- 100000 454 reads randomly pulled out from the total. Only constraint: read length;
- REAS run under default settings;
- Obtained 6325 repeat candidates :
  - Average length: 1186 bp
  - 186 are shorter than 500 bp and/or have depth coverage smaller than 35: artifacts?

Identification (structural search)

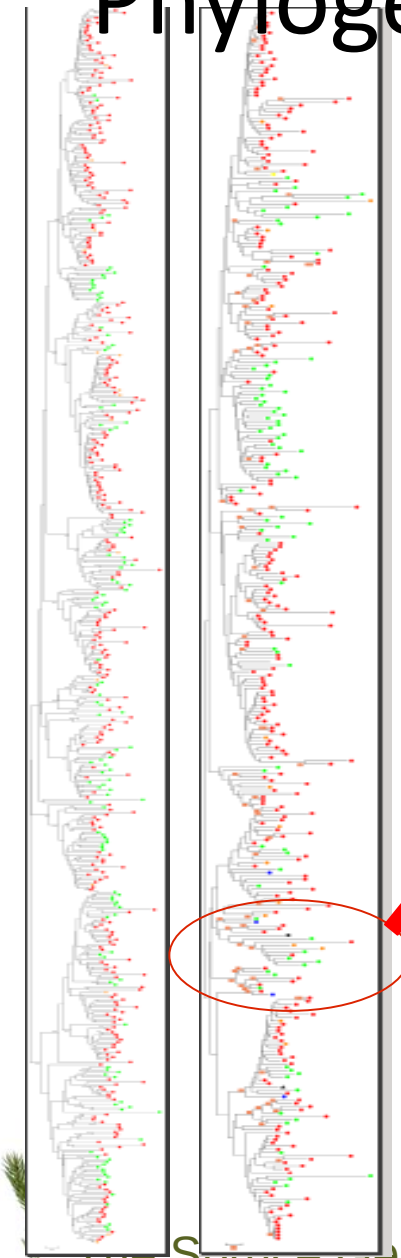


ctgs longer than 10000 nt: were searched using LTR_Finder. The LTR-RT candidates were checked for the presence of TSDs as a guarantee of correct assembly.

Also, positive hits onto contigs were manually checked using dot plot analyses

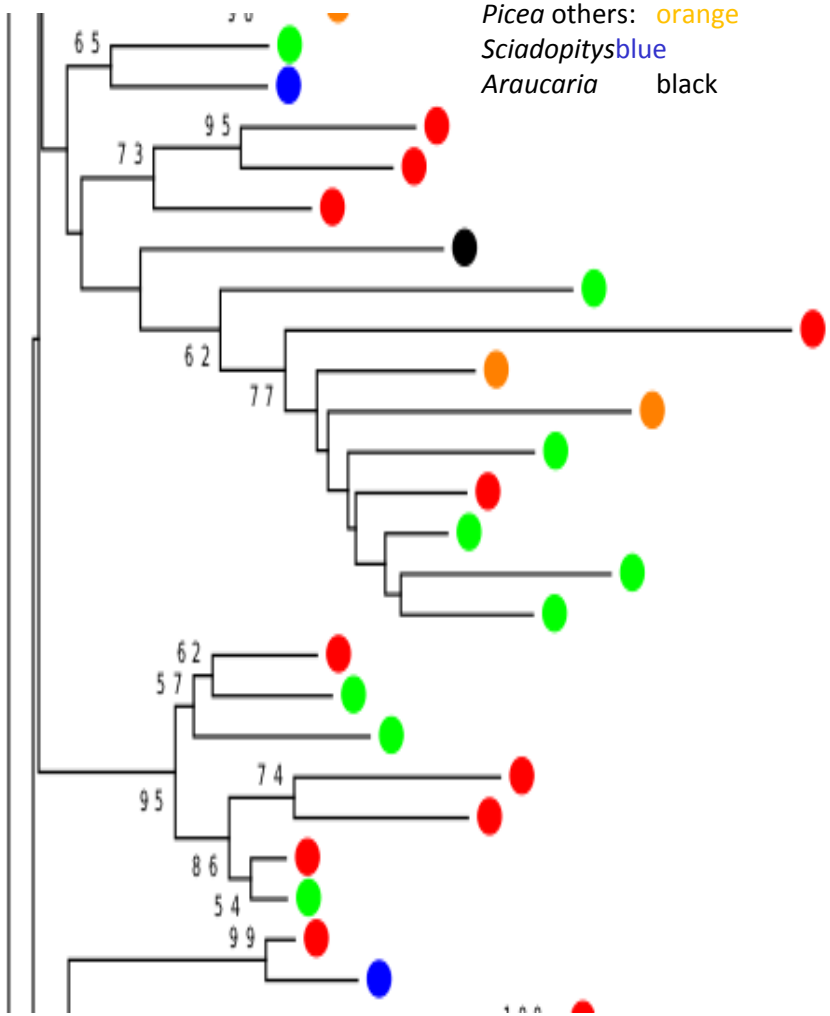
126 complete elements were identified.

Phylogenetic trees



Ty3-gypsy

Ty1-copia



- NJ phylogenetic trees were built using RT domains for both Ty1-copia and Ty3-gypsy elements from *P. abies* and other conifers.
- Ty1-copia showed bootstrap supported clades including elements from different species.