# Functional genomics in Picea  abies

Pär K Ingvarsson
Umeå Plant Science Centre
Umeå University
Sweden

ProCoGen is financially supported by the European Commission under the 7th Framework Programme

Wednesday, February 19, 2014

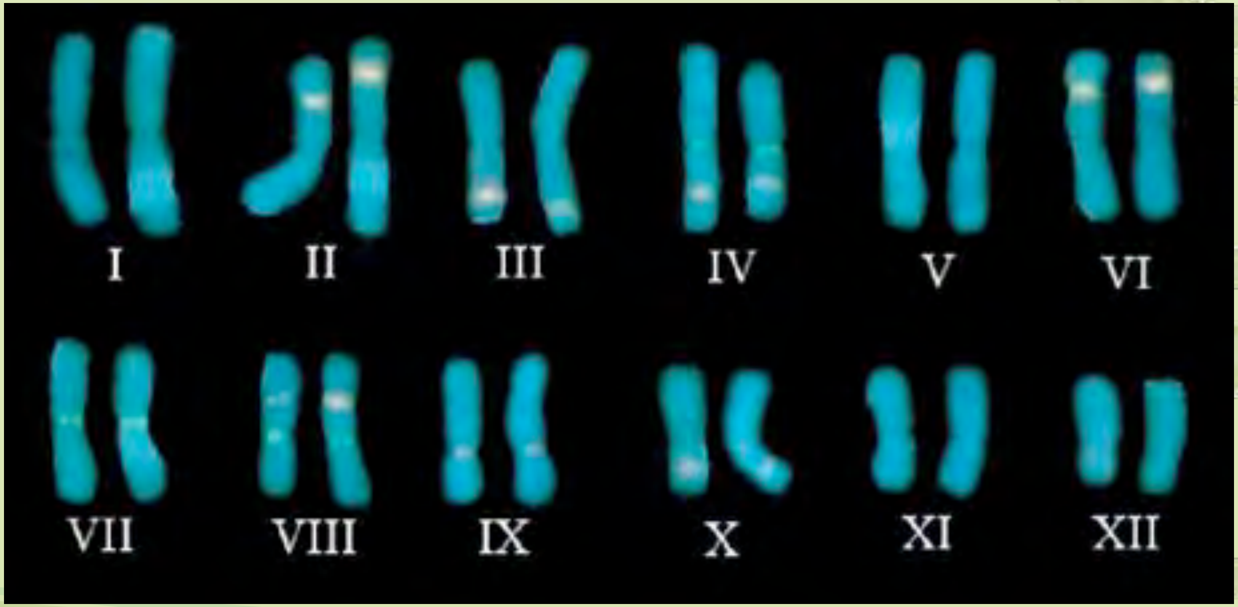
# Scientific and socioeconomic interest

- Science of conifers
  - Evolution: The last major plant group without a sequenced genome
  - Ecology: Dominant members of boreal forests
  - Biology: Unique biological features
- Available genome sequence will aid and facilitate
  - Genomic improvement for biomass productivity, quality, health
  - Optimise cellulose and wood fibre qualities (new materials)
  - Optimised feedstock for bio-refineries
- Norway spruce is Sweden’s most economically important tree
  - 30% net exports
  - 3,000 spruce trees per citizen
  - Annual growth increment worth 2.2 billion €  or ~240 € per citizen per year

ProCoGen is financially supported by the European Commission’s 7th Framework Programme

# Sequencing and assembly

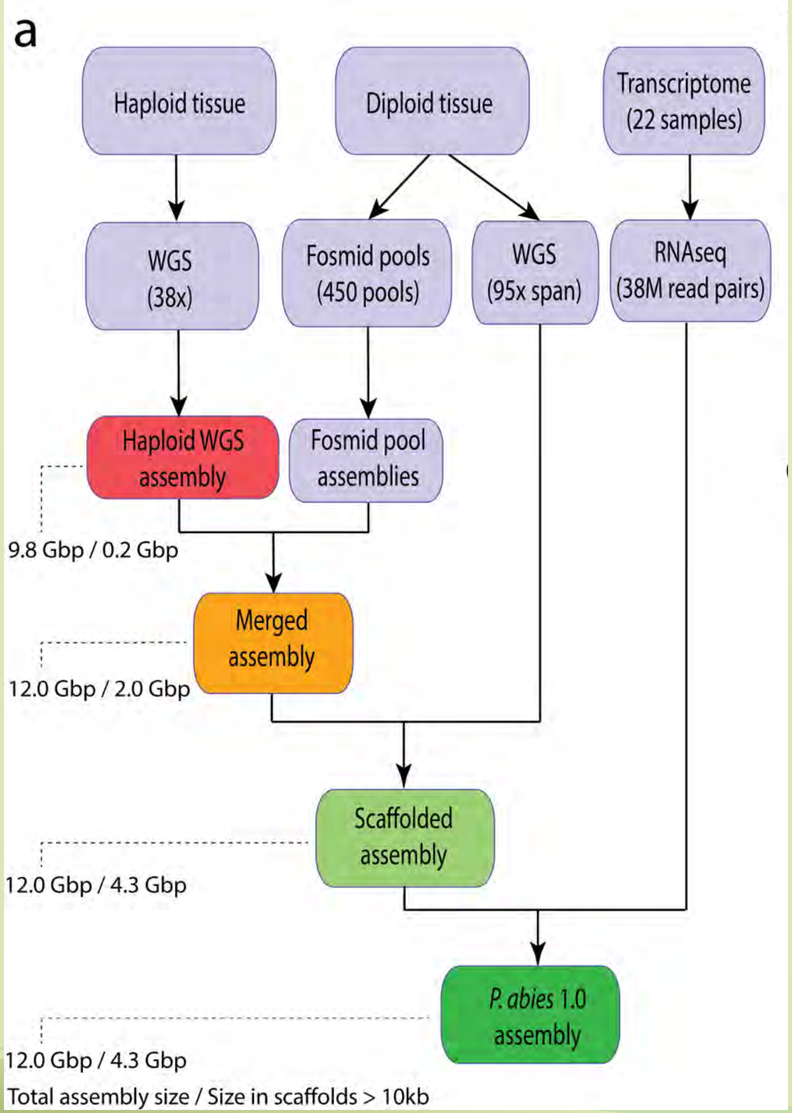
- Challenges
  - 19.6 Gbp genome
  - 12 evenly sized chromosomes (Chromosome sorting not possible)
  - Fairly high heterozygosity
  - High repeat content

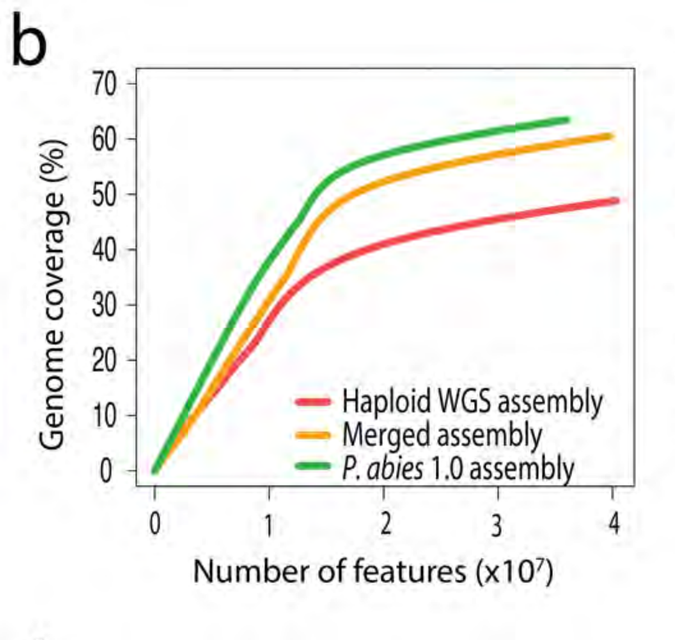


Vischi et al (2003)

ProCoGen is financially supported by the European Commission's 7th Framework Programme

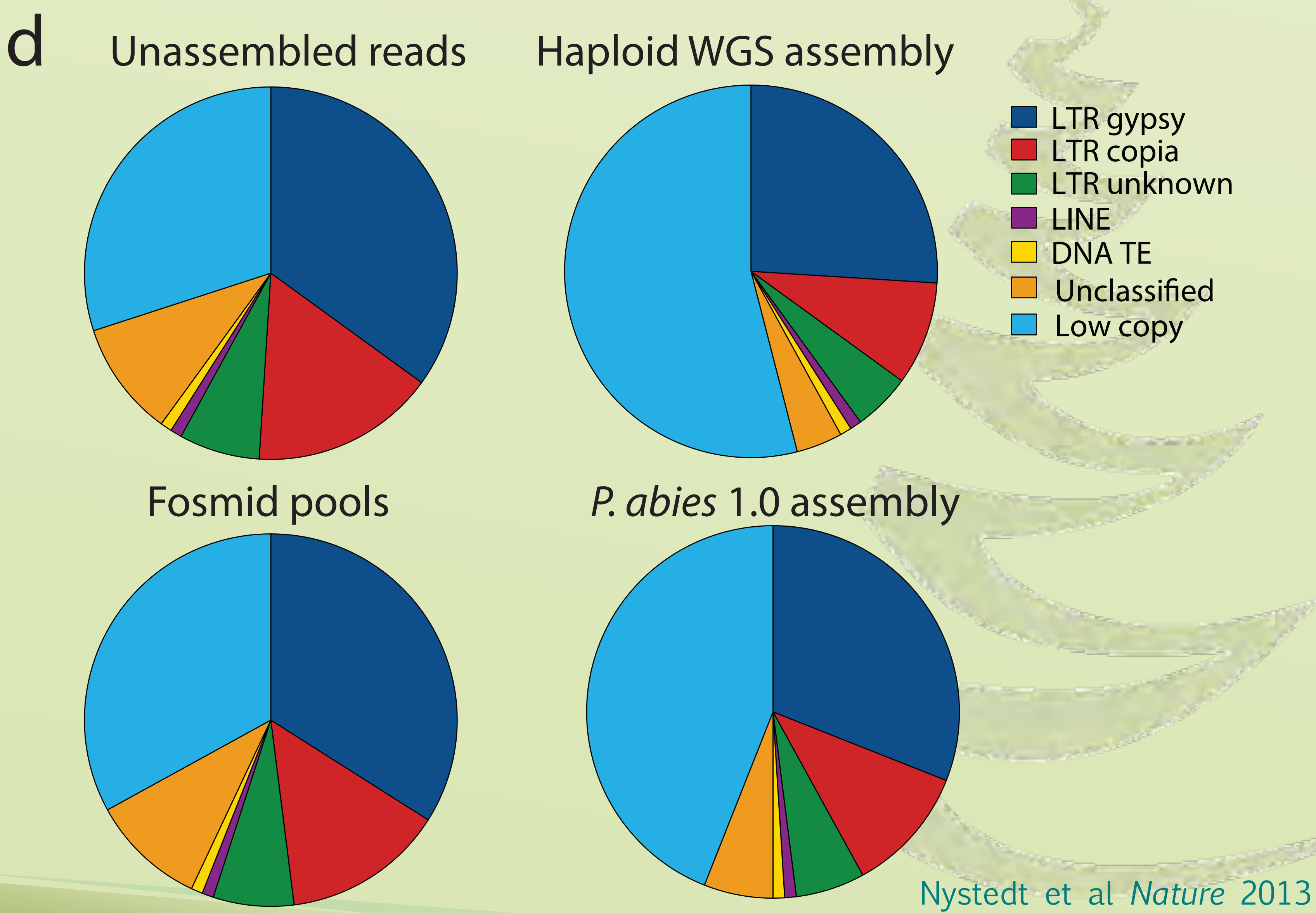
# Hierarchical assembly strategy





ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Fosmid pool sequencing

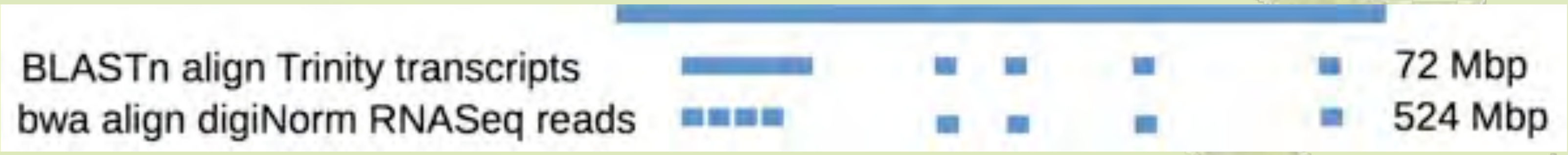


Nystedt et al *Nature* 2013

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Genome sub-selection for gene annotation

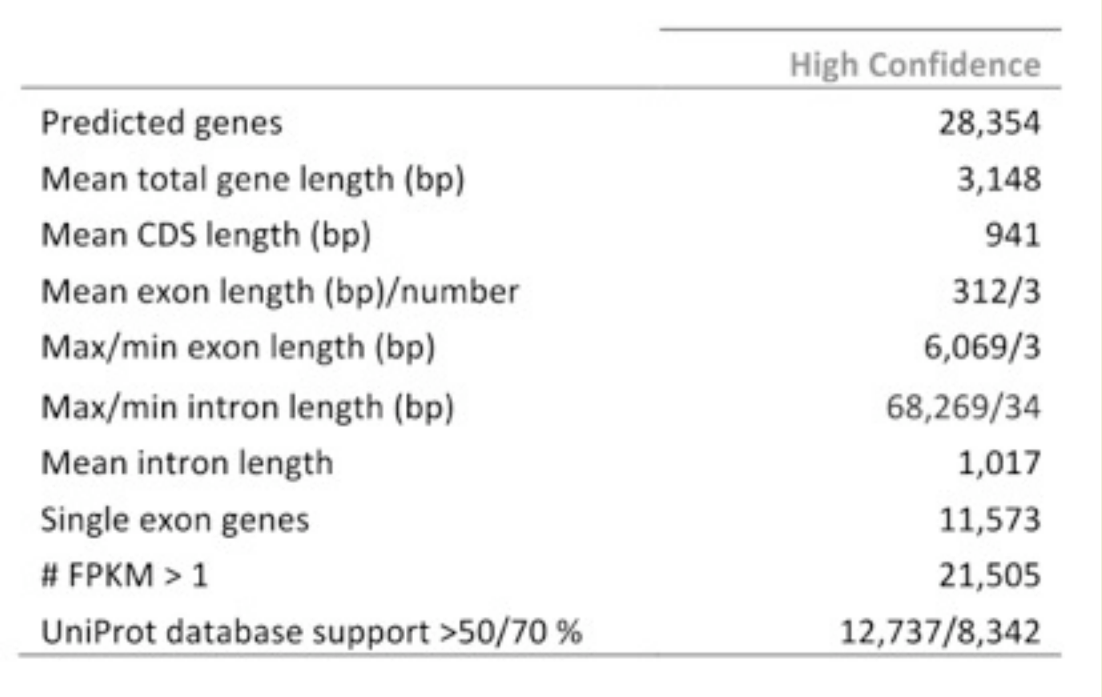
- Sub-selection of genomic scaffolds were done before gene prediction


BLASTn align Trinity transcripts
72 Mbp
bwa align digiNorm RNASeq reads
524 Mbp

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Picea abies gene annotation

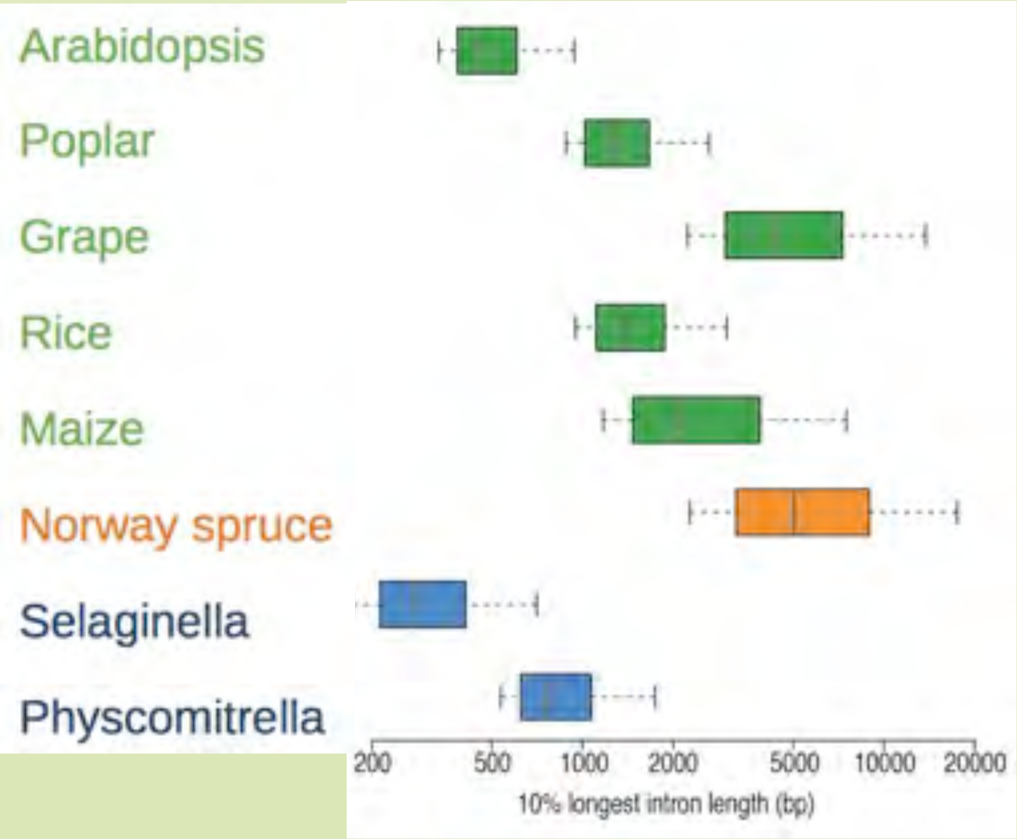
- Augustus+Eugene were run on scaffolds showing evidence for transcription (bwa mapping RNAseq reads) + all Scaffolds >10 Kbp
- 26,597 High Confidence gene predictions
  - >70 % supporting evidence coverage
- 32,263 Medium Confidence predictions
  - 30-70 % coverage
- 8,197 Low Confidence predictions
  - <30 % coverage



| | High Confidence |
|---|---|
| Predicted genes | 28,354 |
| Mean total gene length (bp) | 3,148 |
| Mean CDS length (bp) | 941 |
| Mean exon length (bp)/number | 312/3 |
| Max/min exon length (bp) | 6,069/3 |
| Max/min intron length (bp) | 68,269/34 |
| Mean intron length | 1,017 |
| Single exon genes | 11,573 |
| # FPKM > 1 | 21,505 |
| UniProt database support >50/70 % | 12,737/8,342 |

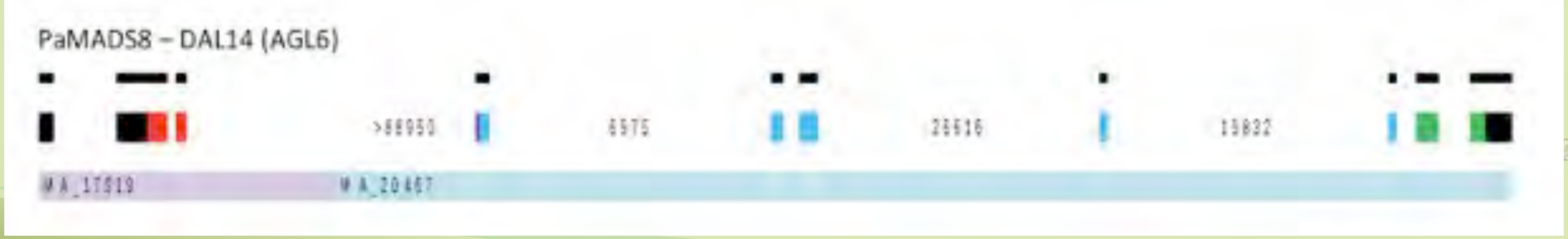
ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Long introns



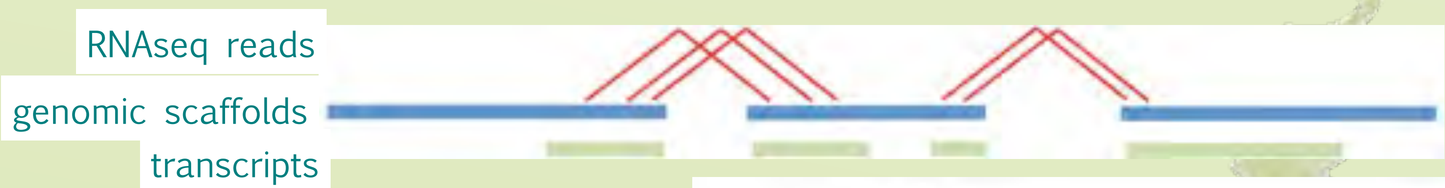
2,384 HC genes contain 2,697 introns >5 Kbp

2,679 contain TE



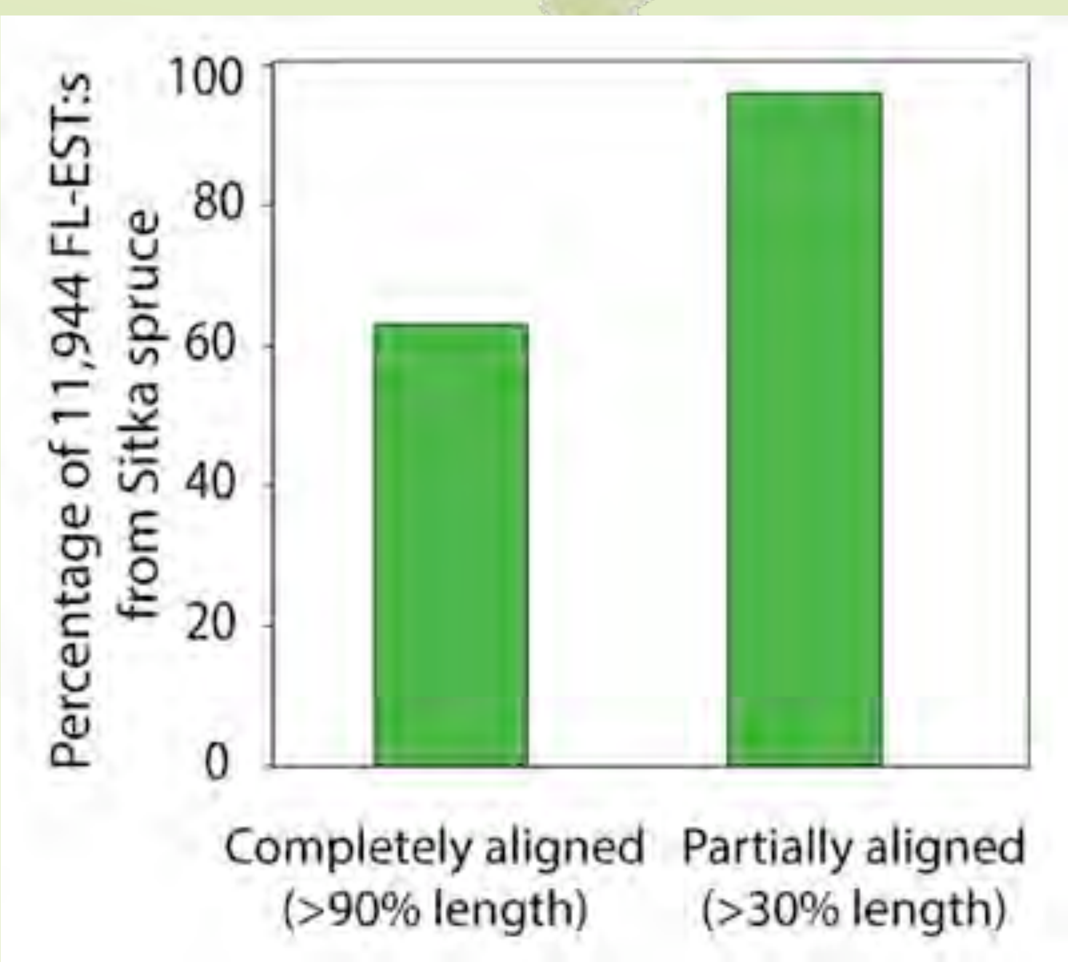
ProCoGen is financially supported by the European Commission's 7th Framework Programme

# RNAseq scaffolding



- For a relatively large fraction of genes, assembly breaks in long (>10kb) introns
- Scaffolding using RNAseq data alleviates this to some extent (59% → 64%)

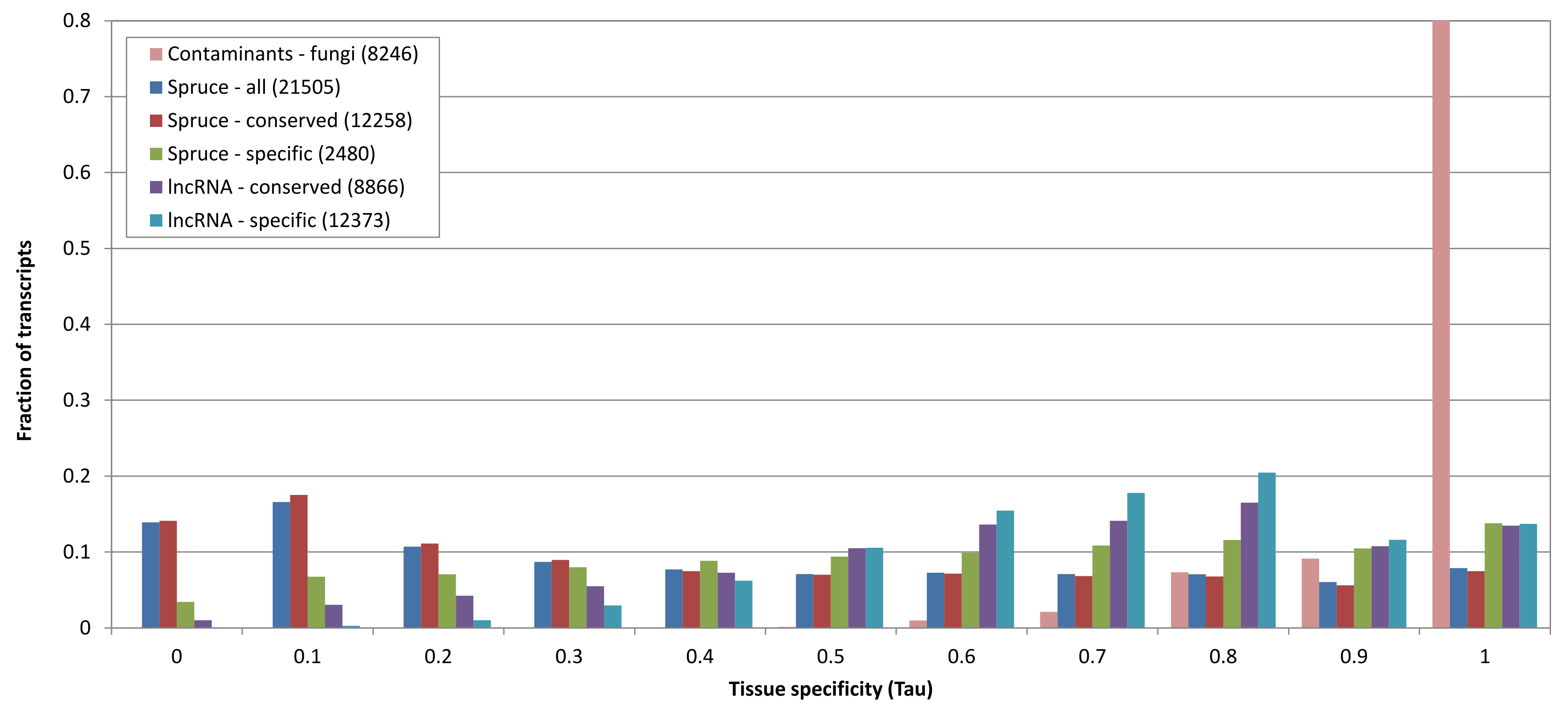
- Transcript, PacBio and gene fusion based scaffolding
- expected to complete ~5000 additional genes



Nystedt et al *Nature* 2013

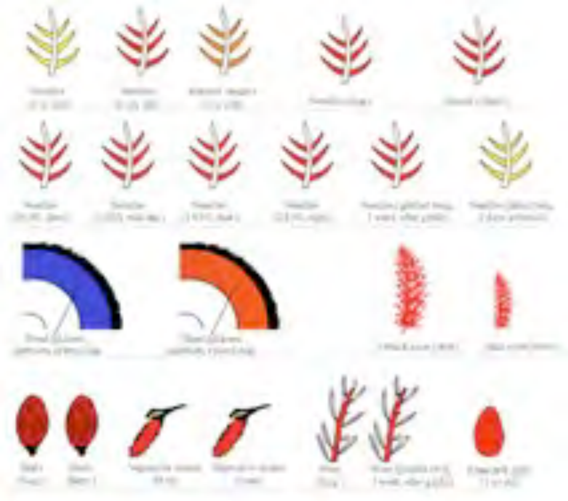
ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Expression across tissues

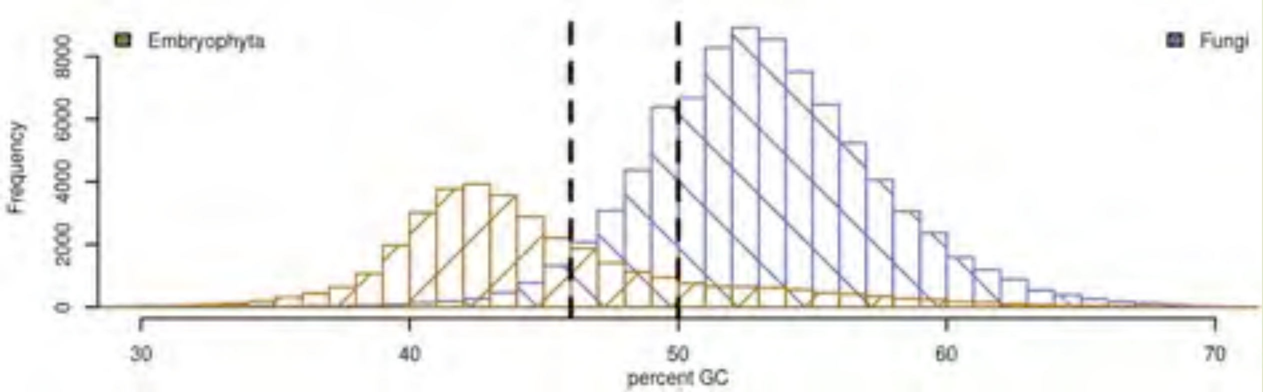


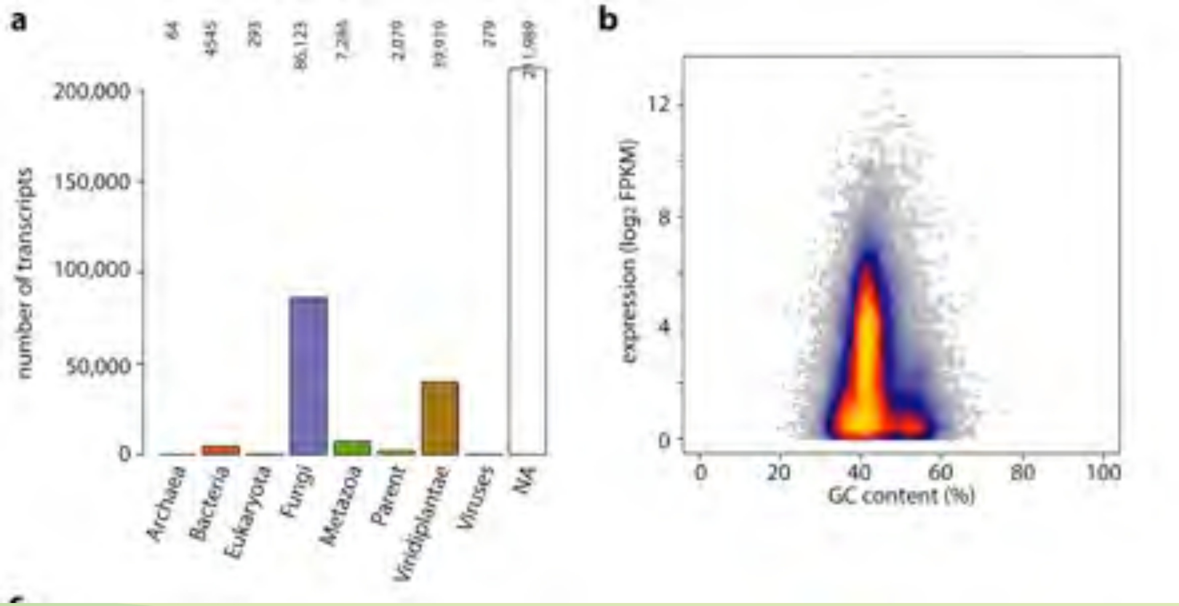
**ProCoGen is financially supported by the European Commission's 7th Framework Programme**

# An expression catalogue



22 libraries, 50 M PE reads per library





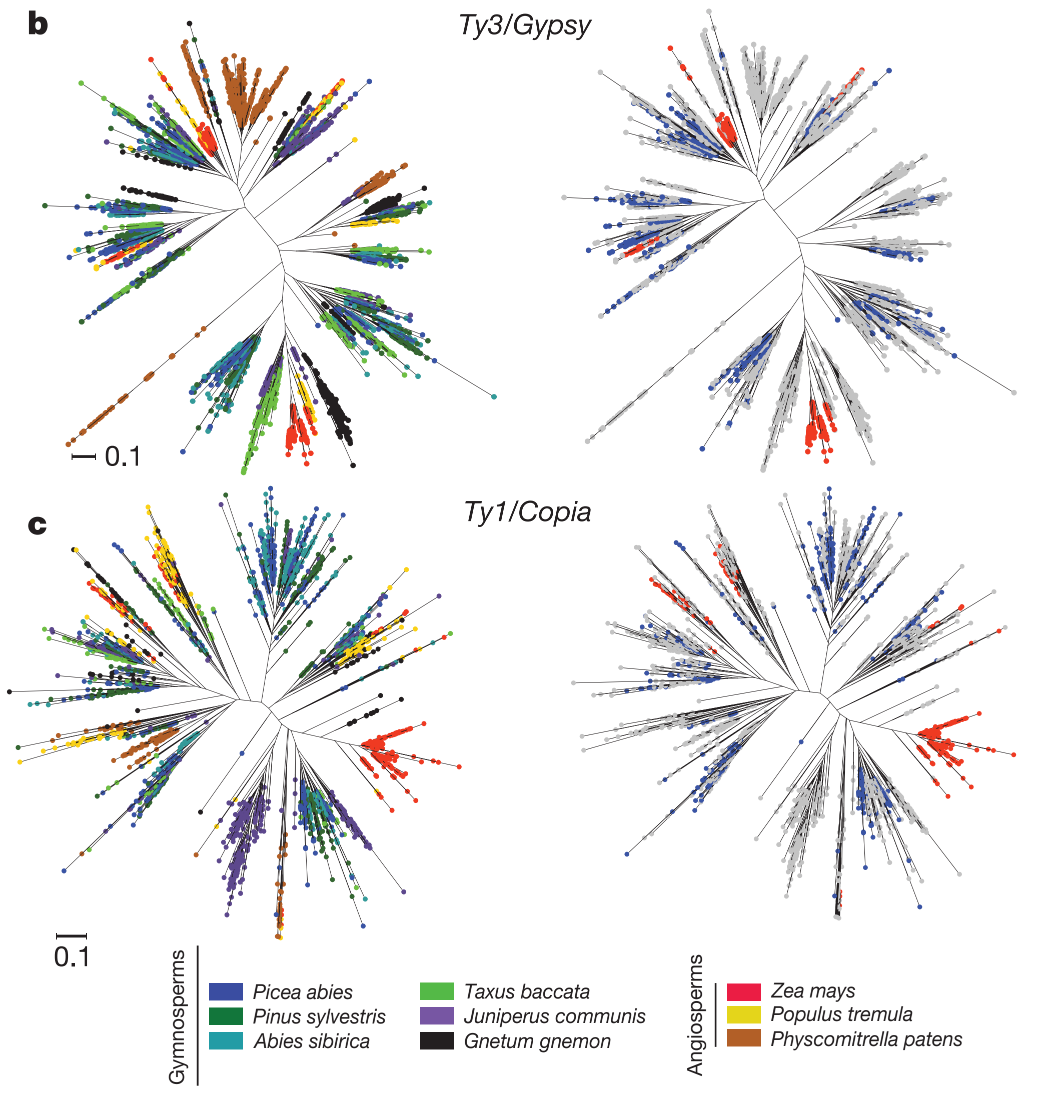
ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Identification of repetitive sequences

- Identitification
  - For each species: random samples of 100000 454 reads longer than 500 bp
  - RepeatScout + cap3 + cd-hit-est clustering

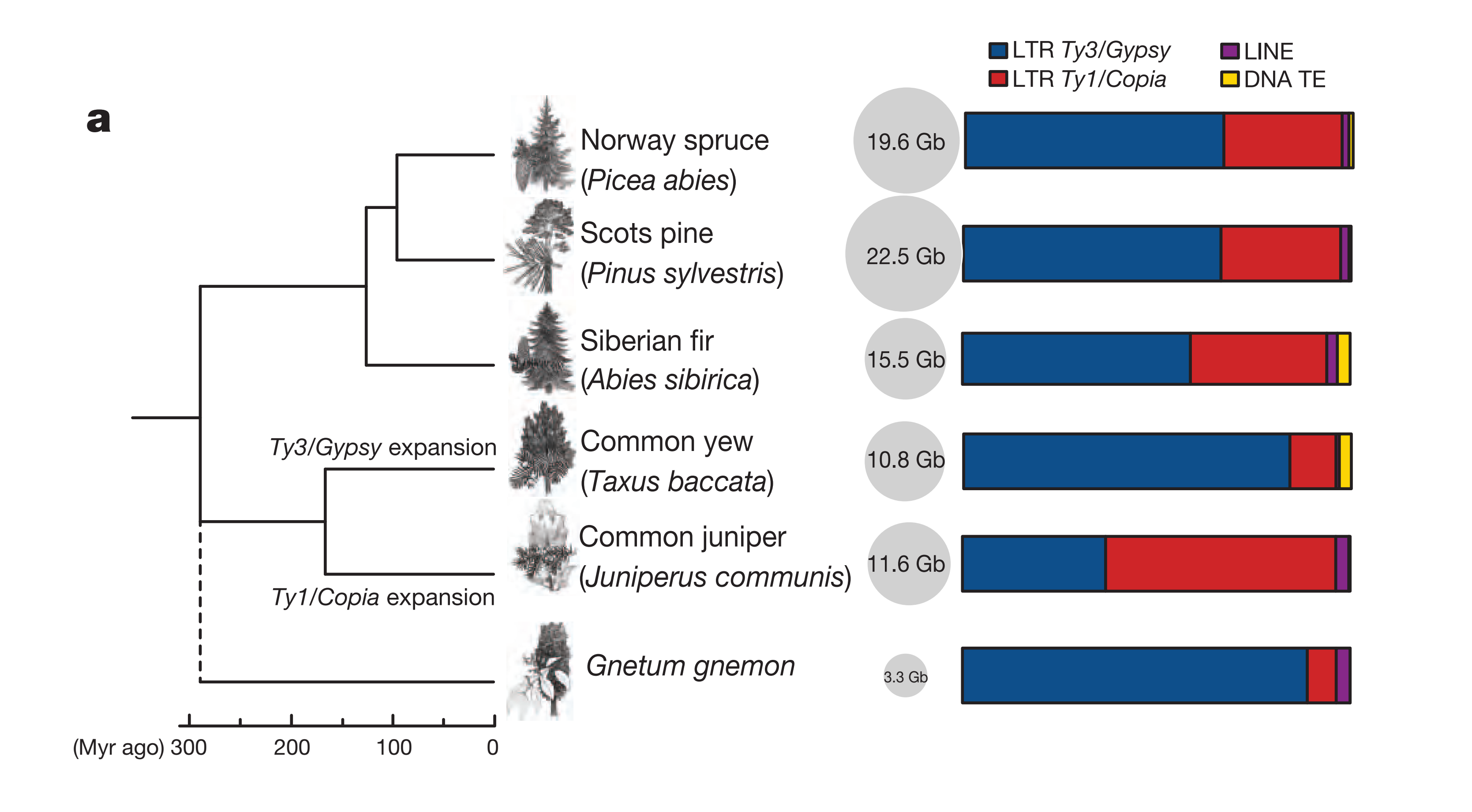
- Characterization
  - Similarity (blastX and tblastx) searches against other TEs database and nr GenBank.
  - removal of plastid contaminants and candidates having hits with known gene families

# LTRs are widespread ad shared between species



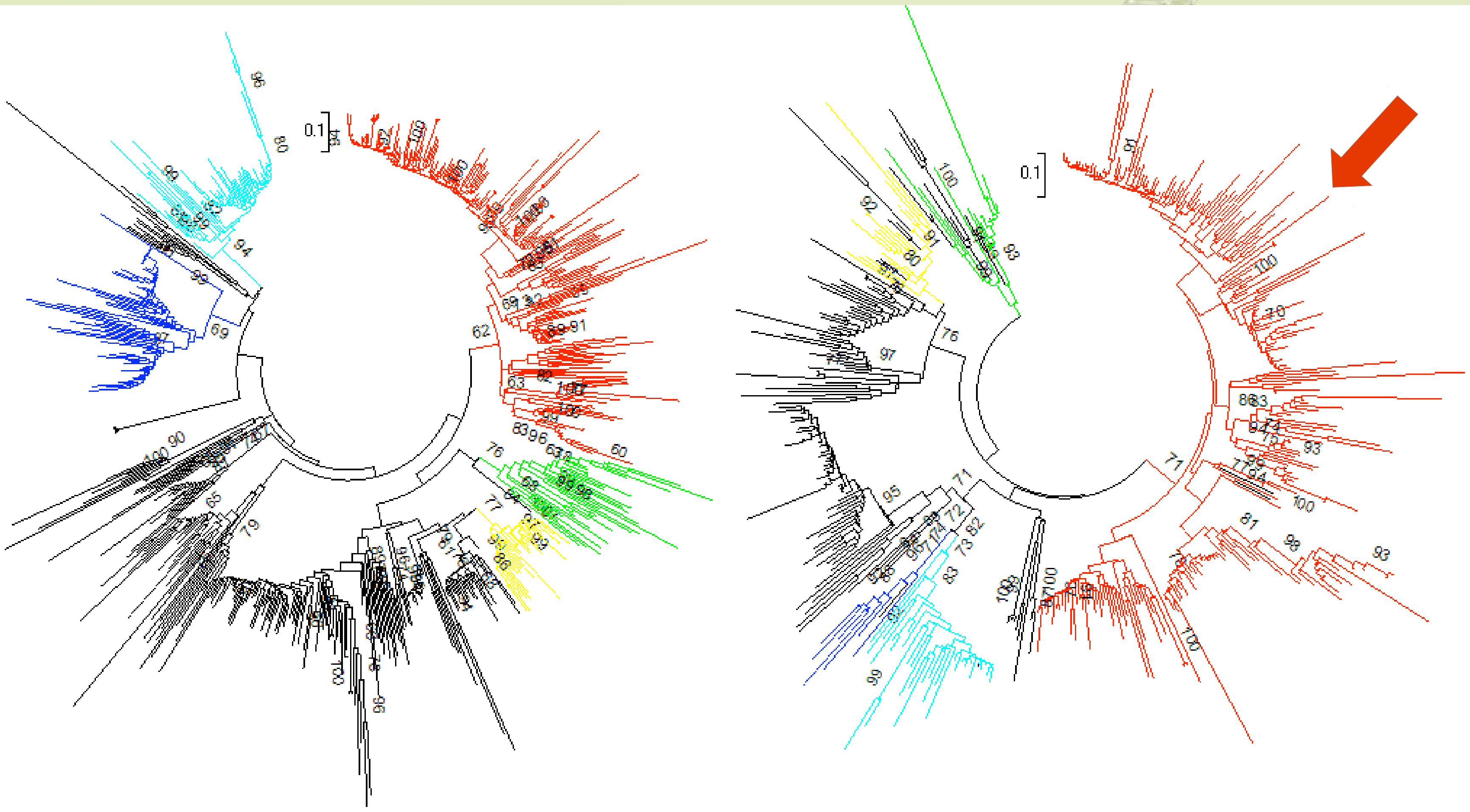
ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Repetitive structure of the Norway spruce genome



ProCoGen is financially supported by the European Commission's 7th Framework Programme

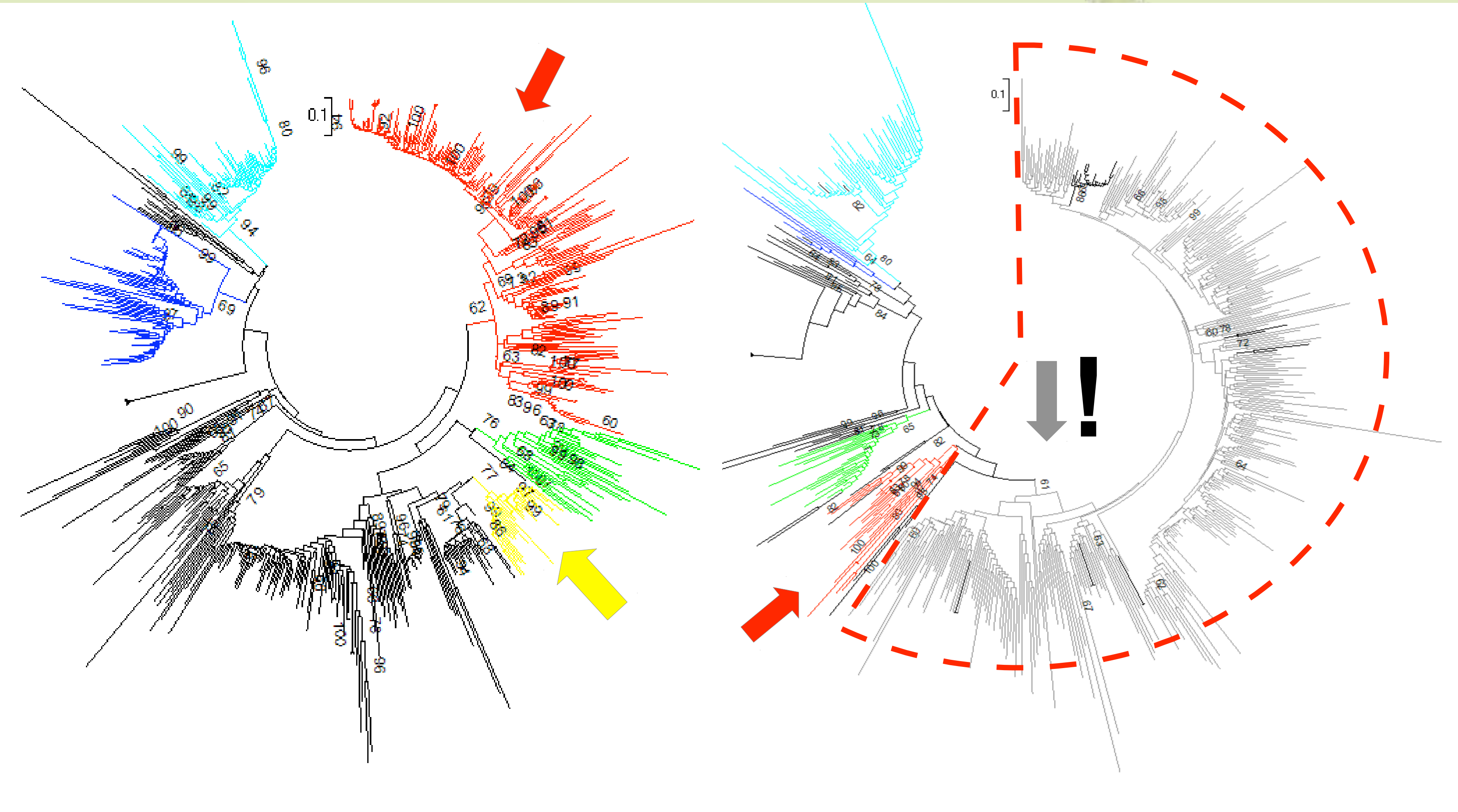
# The same LTR group have different fates in different conifer species



*P. abies*  Ty1-copia  *Abies alba*

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# The same LTR group have different fates in different conifer species - an extreme case!



*P. abies*

Ty1-copia

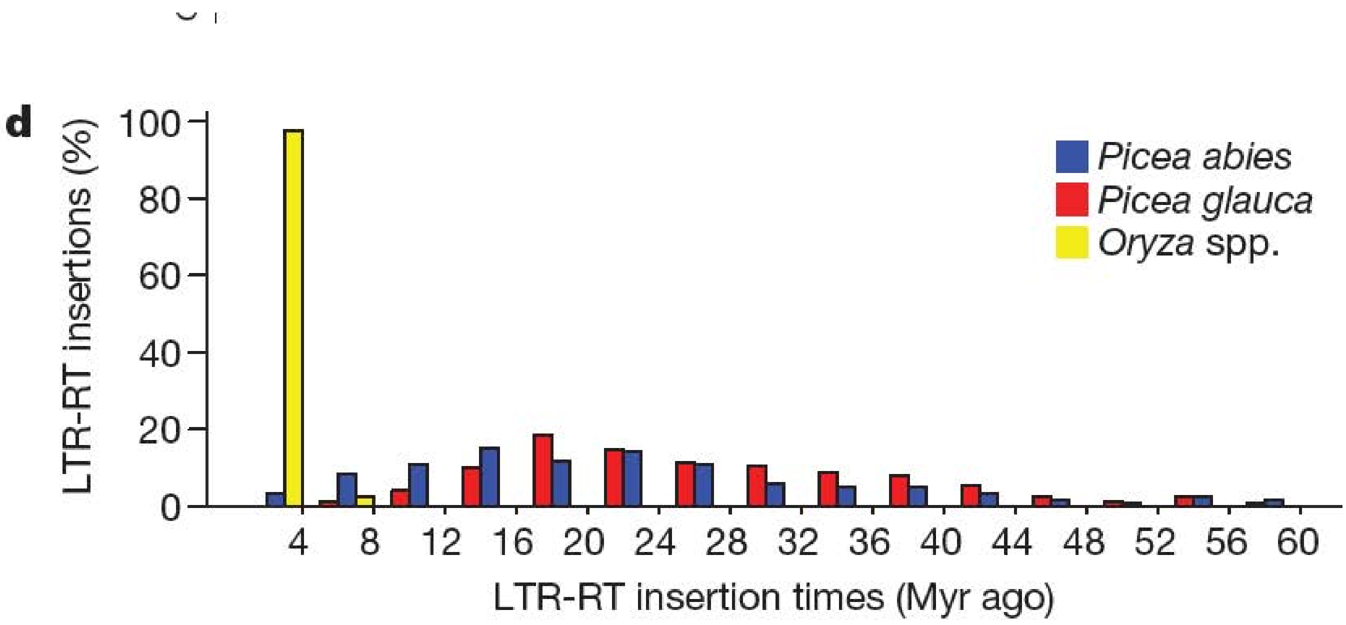
*Juniperus communis*

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Summary of the repetitive fraction

- Conifer large genomes are mostly repetitive
- LTR-RTs are the most represented TE class and are shared across different genera;
- LTR-RTs from the same group had different fates in different species...
- ...nothing unexpected and nothing that different from what was already described in several angiosperms!

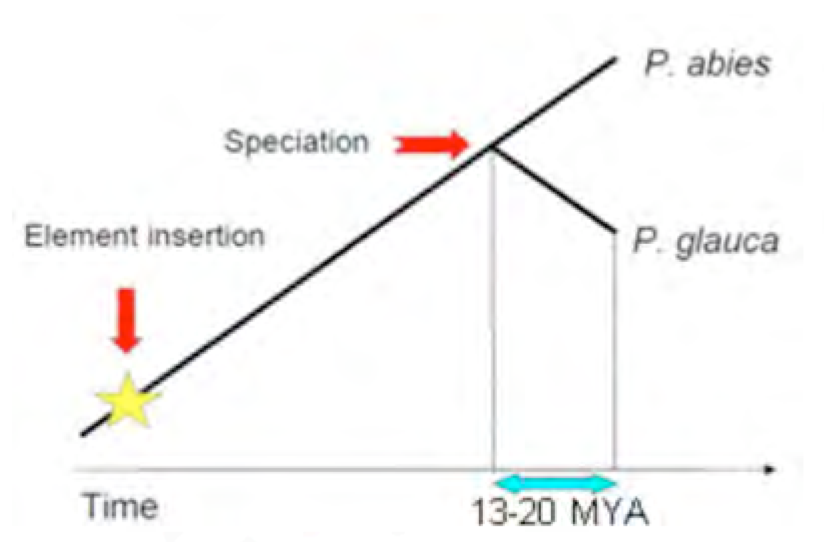
# Amplification of conifer LTRs occurred in ancient times

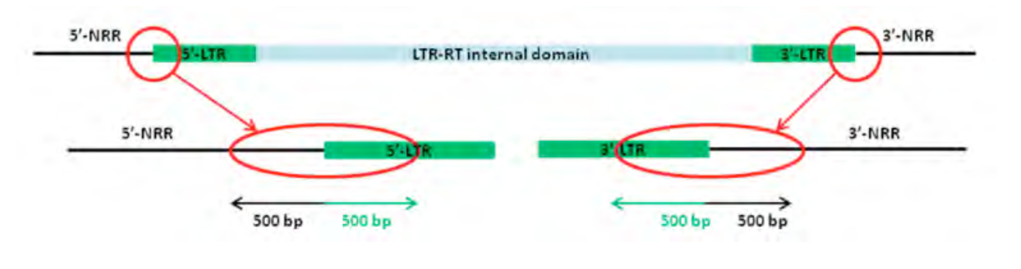


ProCoGen is financially supported by the European Commission's 7th Framework Programme

# LTR amplification in Picea mostly pre-dates speciation

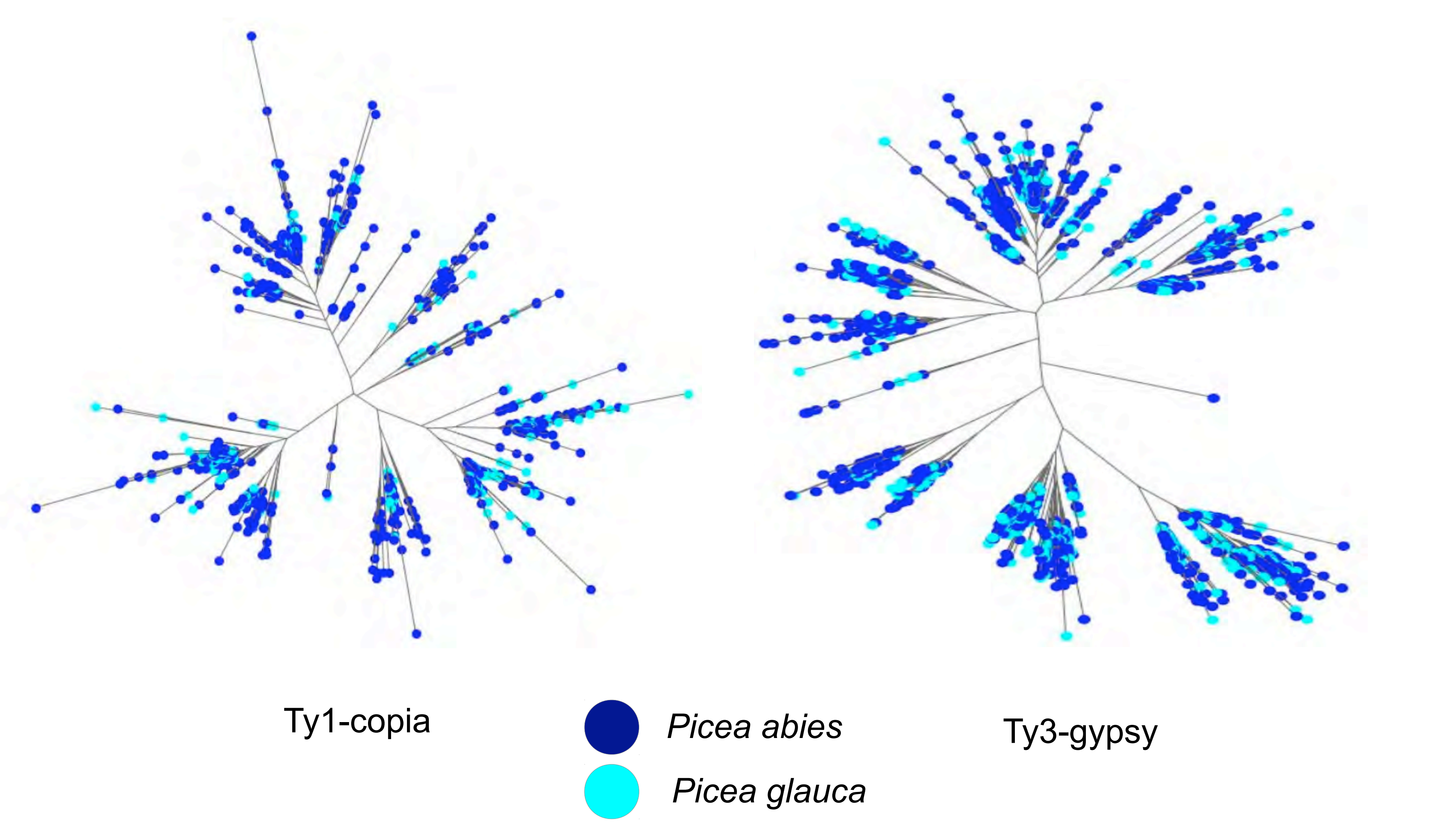
- 113 LTR-RTs complete elements were identified in 25 *P. glauca* BACs
- The tracts spanning for 1000 bp all the boundaries genome-LTR-RT-genome were mapped onto *P. abies* assembly
- 45 elements: not mappable
- 5 inserted in *P. glauca* after speciation (empty spot in *P. abies*); younger than 13 MY
- 63 inserted before speciation: older than 13 MY





ProCoGen is financially supported by the European Commission's 7th Framework Programme

# LTR amplification in Picea mostly pre-dates speciation



ProCoGen is financially supported by the European Commission's 7th Framework Programme

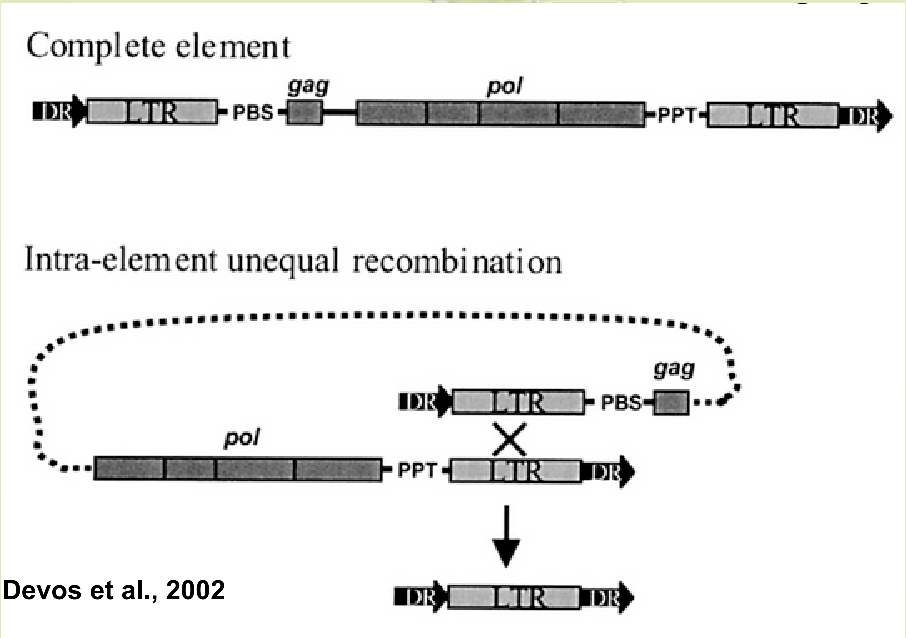
# Summary of LTR insertion dynamics

- LTR-RTs massively amplified in ancient times;
- There is no (or limited) evidence of recent LTR-RTs insertional activity;
- A huge amount of ancient inserted LTR-RTs escaped removal
- The emerging scenario is totally different from the one in angiosperms and leads to a crucial question:
- Why (How) these elements have been retained?

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# LRT removal through unequal recombination

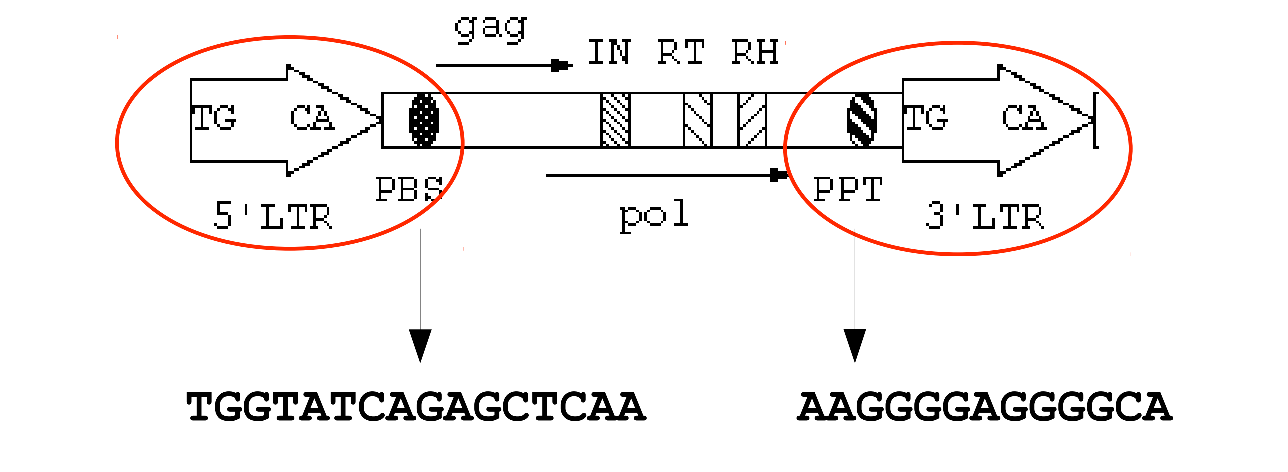
- We searched for evidence of UR for 3 LTR-RT families in *P. abies*
- We used 4 BACs, 20 fosmids and the assembly contigs longer than 50 kbp



Devos et al., 2002

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Estimating LTR removal through unequal recombintion



- LTRs can be assigned to complete elements or identified as a solo-LTR
  - If it is followed by PBS...it's the 5' LTR of a complete element;
  - If it is preceded by PPT...it's the 3' LTR of a complete element;
- If there is no evidence of PPS or PPT and there are TSD at end → solo LTR
- Any other case → UNCLEAR (amount should be negligible)
- Ratio complete_elements: sLTRs ---> (5'LTR +3'LTR)/2 : sLTR

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Estimating LTR removal through unequal recombintion

| LTR Family | Complete elements | Solo LTRs | Ratio |
|---|---|---|---|
| ALISEI | 27 | 5 | 0.185 |
| 3K05 | 26 | 5 | 0.192 |
| 4D08_5 | 43 | 0 | 0.0 |
| ***Total*** | ***96*** | ***10*** | ***0.104*** |

- In *P. abies* the ratio of solo-LTRs to complete elements is ~1:9
- In *A. thaliana*, rice and barley the corresponding ratios are 1:1, 0.6:1 and 16:1.

**ProCoGen is financially supported by the European Commission's 7th Framework Programme**

# The emerging picture: how did conifer genomes become so large?

Massive LTR-RTs amplification
occurred for a long period in ancient
evolutionary times

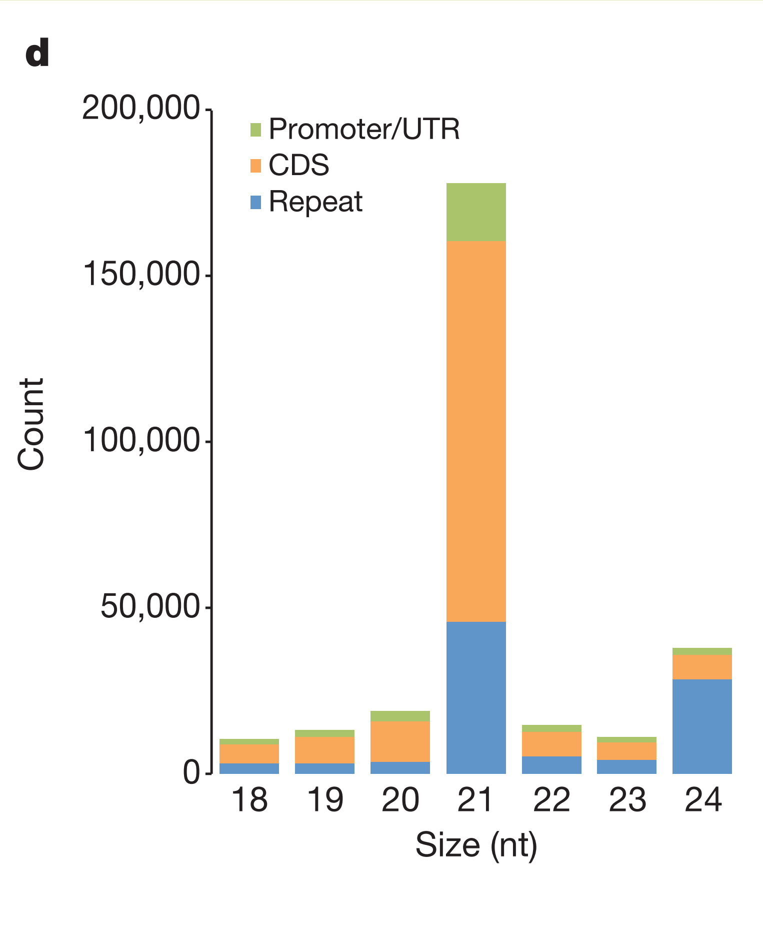
+

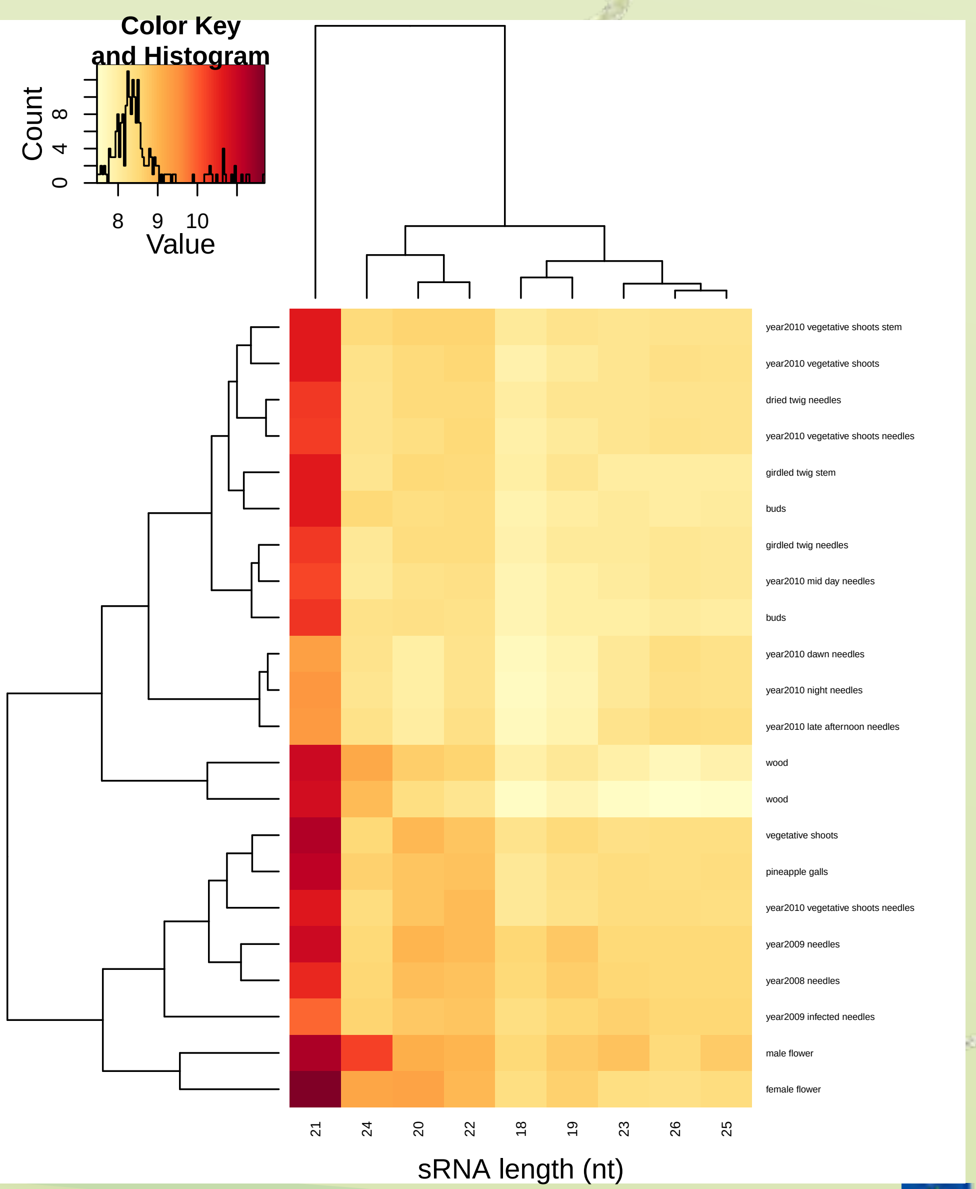
(very) limited homologous recombination
targeting these elements
– No LTR-RTs removal;
– No LTR-RT mediated damage to the
host genome

=

LTR-RT accumulation → Large genomes

# sRNA distribution





ProCoGen is financially supported by the European Commission's 7th Framework Programme

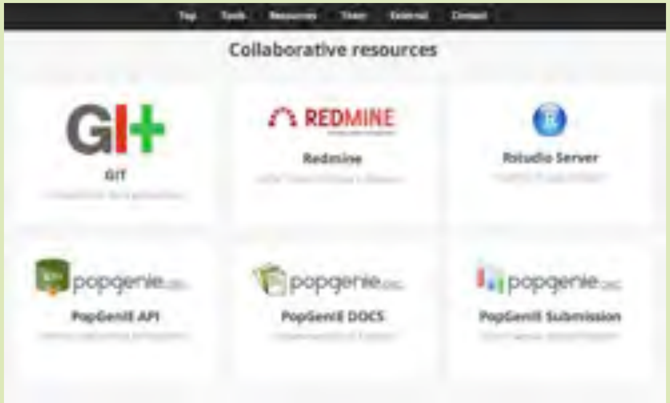
# In-progress RNAseq and sRNA studies

- Root, needle, cambium and soil seasonal profiling: control and long-term fertilized + metagenomics and metatranscriptomics
- Needle developmental profiling
- Phloem and xylem seasonal profiling
- Developing wood cross section
- Somatic embryogenesis
- Reproductive organ development
- Embryo and seedling development
- Suspensor cell
- Diurnal expression in needles and cambium

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# Bioinformatic resources

PlantGenIE.org



popgenie
Welcome to v3 PopGenIE: The Populus Genome Integrative Explorer

ProCoGen is financially supported by the European Commission's 7th Framework Programme

# The Norway 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

**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

**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

**SAB**
Kerstin Lindblad-Toh
John MacKay
Outi Savolainen
Detlef Weigel

**Skogforsk**
Bengt Andersson
Bo Karlsson

**SNIC Supercomputers**
Uppmax/PDC/NSC/HPC2N

**SNISS national infrastructure**

**CLCbio**

**Lucigen**

Knut och Alice Wallenbergs Stiftelse

Umeå Plant Science Centre
a centre of excellence

SciLifeLab

IGA

VIB

Umeå Universitet

SLU

Stockholm University

Karolinska Institutet

Uppsala Universitet

C·H·O·R·I

SKOGFORSK