A first look at the large and complex genome of Norway spruce (*Picea abies*)

Pär K. Ingvarsson

The Spruce genome project
Umeå Plant Science Centre
Department of Ecology and Environmental Science
Umeå University
SE-901 87, Umeå Sweden

The Spruce Genome Project

Why sequence the Norway spruce genome?

The scientific argument:

- Evolutionary interesting group
  - conifers are evolutionary ancient and the last major plant group which have not at least one member with a complete genome sequence
- Ecological importance
  - conifers are dominant members of many ecosystems, primarily in boreal forests
- Unique biology?

Why sequence the Norway spruce genome?

The strategic argument:

- Norway spruce is the economically most important Swedish tree
- Genome sequence will spark research to generate:
  - New tools for breeding for tree productivity, quality, health
  - New knowledge and tools for cellulose and wood fibre modification (new materials)
  - New knowledge and tools for tree-based biorefineries

Challenges with sequencing a conifer genome

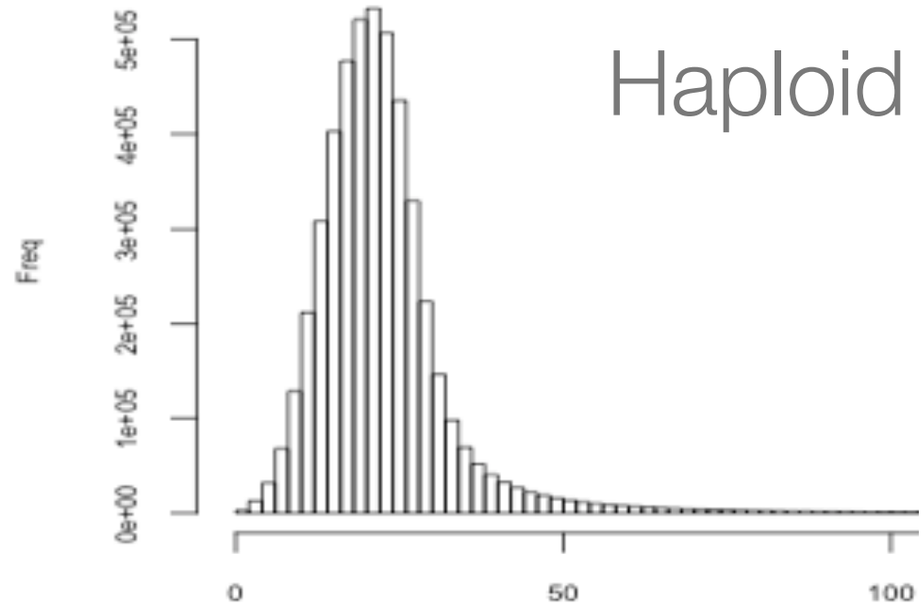
- Huge genome of approximately 20 Gb – seven to ten times the human genome
- The Norway spruce genome will be the largest genome sequenced so far
- >99% is likely moderately or highly repetitive DNA of unknown function
- Large gene families and abundant numbers of pseudo-genes
- <3% consists of sequences homologous to genes.

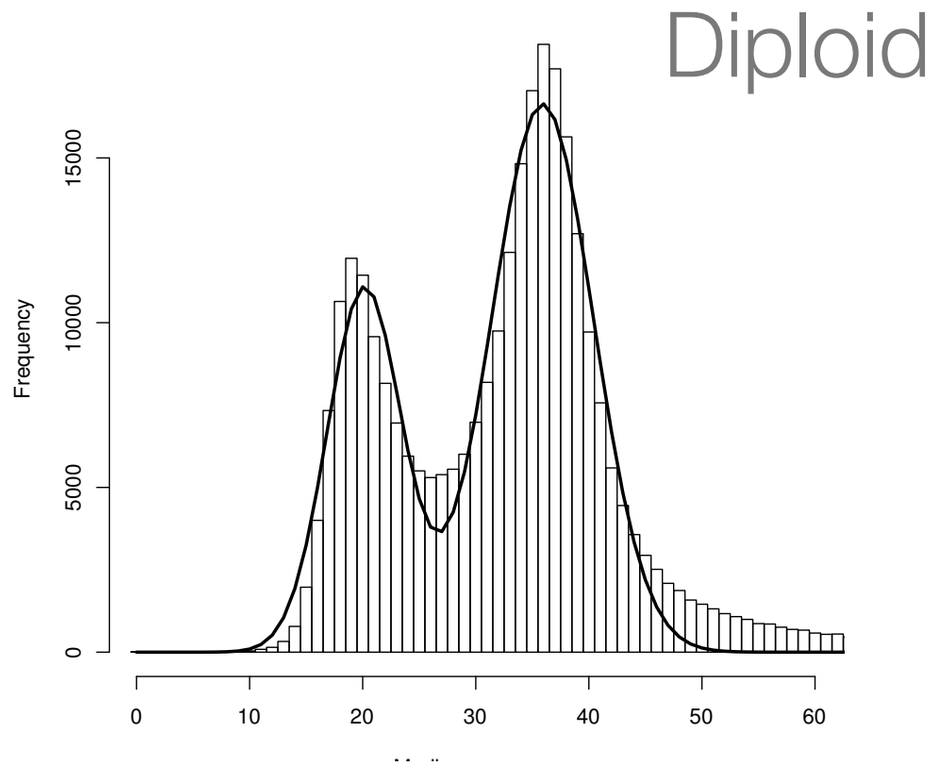
Sequencing status

Single-end Roche 454	1x read coverage ~450 bp 1x read coverage ~650 bp
Paired-end Illumina 2*100 bp	>50X read coverage 150, 300 and 650 bp fragments
Mate-pairs Hybrid Roche/Illumina	30X span cov 3 kb fragments 20X span cov 5 kb fragments
Fosmid ends	20X span cov
Fosmid pools	500 000 40 kb fosmids in pools of 1000
Silver standard Fosmids Roche 454	100 X
Mitochondria Roche 454 (long)	100 X

Current genome assemblies

	Haploid	Diploid
Coverage	20x (10x?)	50x + 1.5x 454
Contigs > 1 kbp	30%	44%
Contigs > 5 kbp	8%	12%
Contigs > 10 kbp	1%	3%
NG50:	204 bp	757bp





The Spruce Genome Project

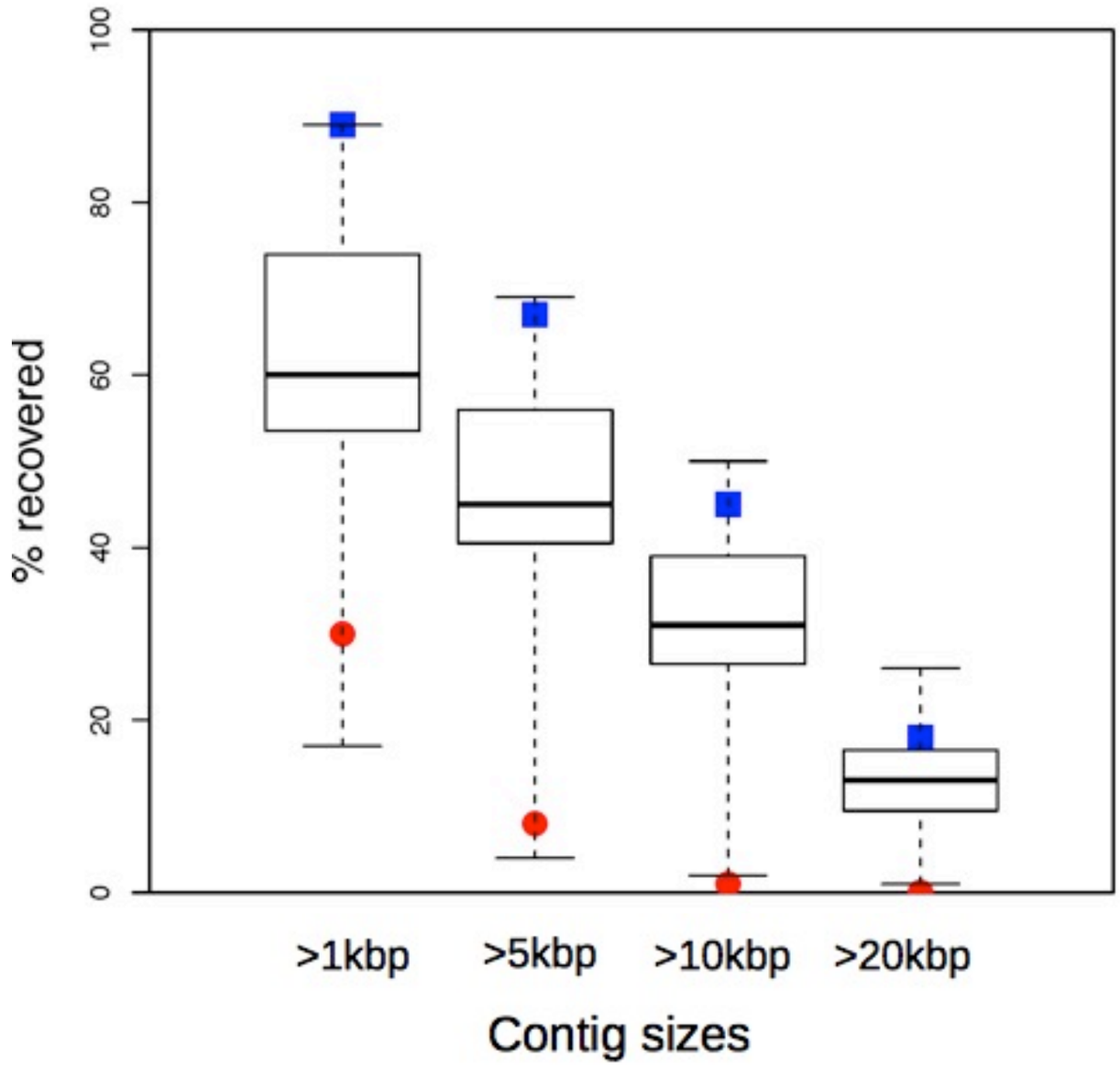
Fosmid pool strategy

Fosmid pools:

First round:
500 pools (1x)
300 libraries made
56 pools sequences and analyzed

Second round:
1500 pools (3x)
Production in progress

Even with relatively few pools analyzed results already “beat” WGS assemblies



More WGS information:

Genome assembly:

- *Sunday Jan 15, 4:10 pm:* Sequencing and assembly of the largest and most complex genome to date - the Norway spruce (*Picea abies*) - *Björn Nystedt*

Fosmid sequencing:

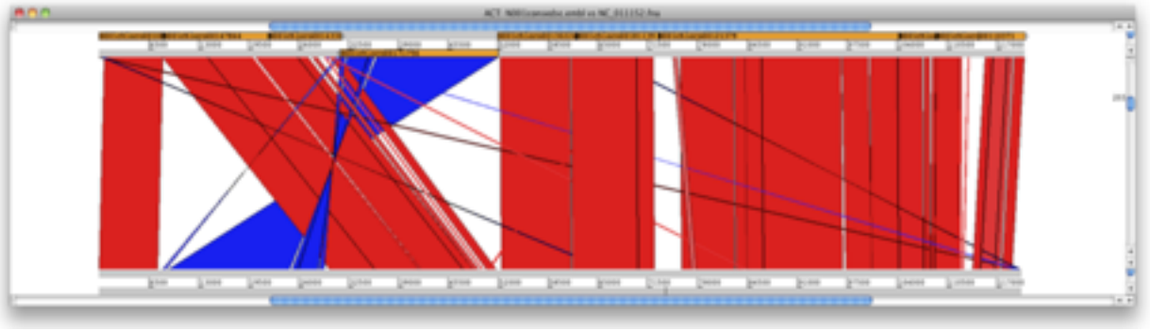
- *Tuesday Jan 17, 2:10 pm:* Fosmid pool sequencing of the 20 Gbp genome of Norway spruce (*Picea abies*) - *Björn Nystedt*

Sequencing the *Populus tremula* genome:

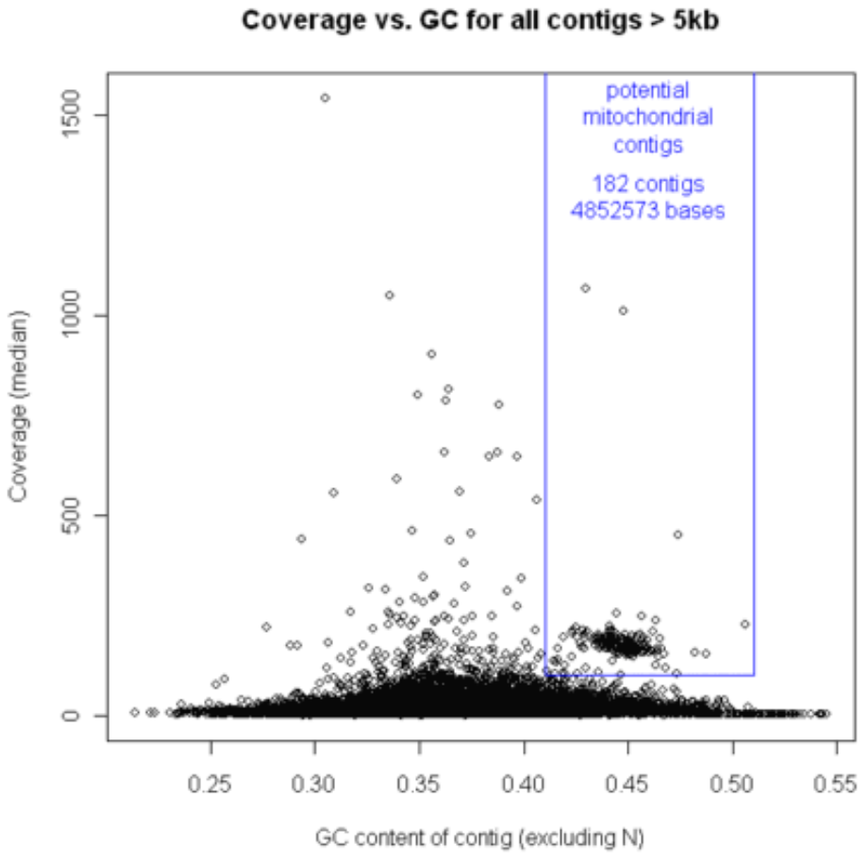
- *Sunday Jan 15, 11:30 am:* The genome of aspen (*Populus tremula*) - the most complex genome sequenced to date? - *Stefan Jansson*

Picea abies organelle genomes

- cpDNA:
  - Chloroplast genome of expected size ~120 kbp
  - 1 run 454 assembled into 9 chloroplast contigs
  - Scaffolding with 1% of 1 lane of Illumina MP data => 1 circular scaffold
  - Detected 1 translocated inversion compared to *Picea sitchensis*

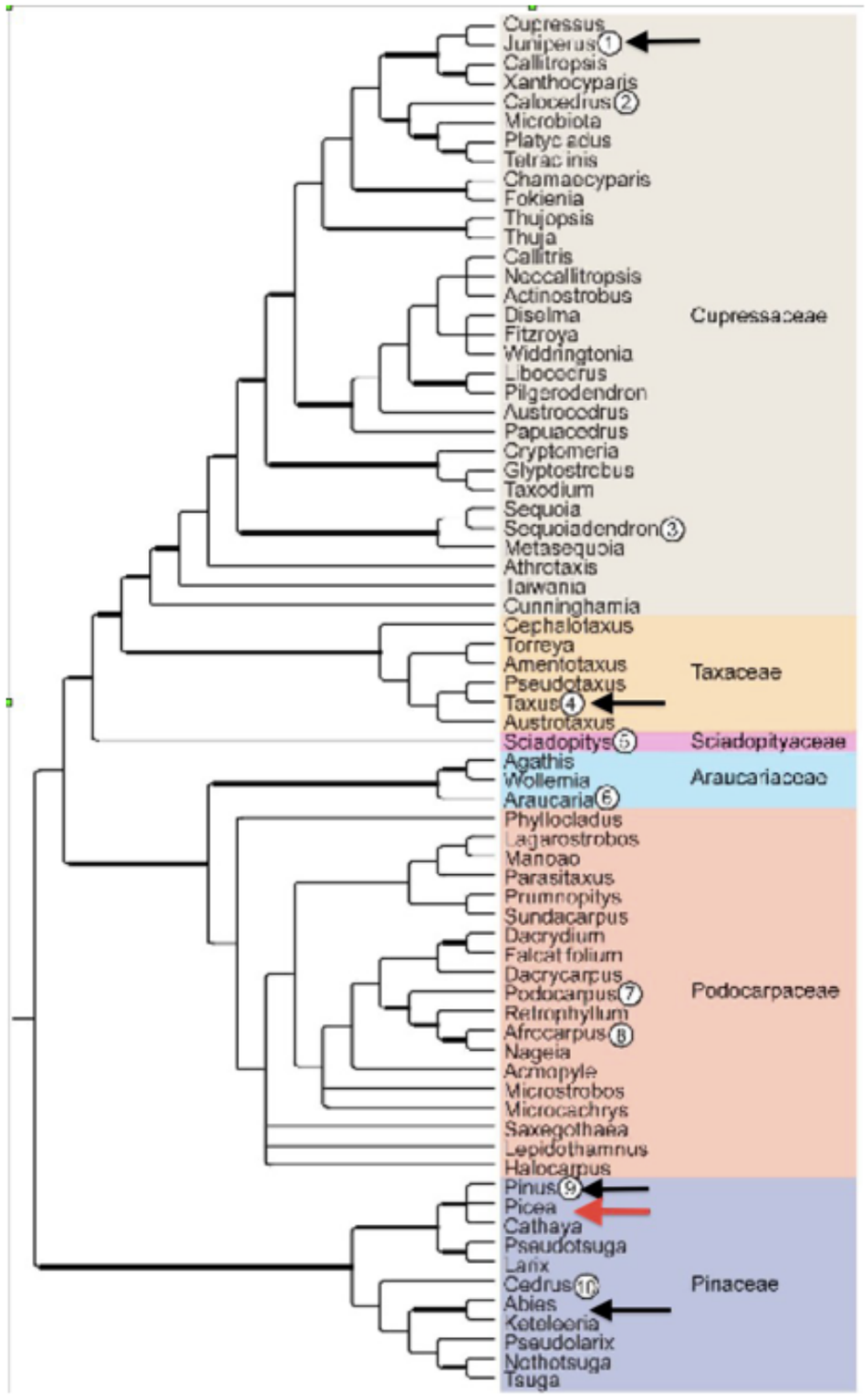


- **mtDNA:**
  - Mitochondrial genome possibly large - estimated size ~4.8 Mb
  - N50 contigs: 50 kbp
    N50 scaffolds: 289 kbp
  - Much longer contigs than nuclear DNA - less repeats?



Comparative genomics of conifer repetitive elements

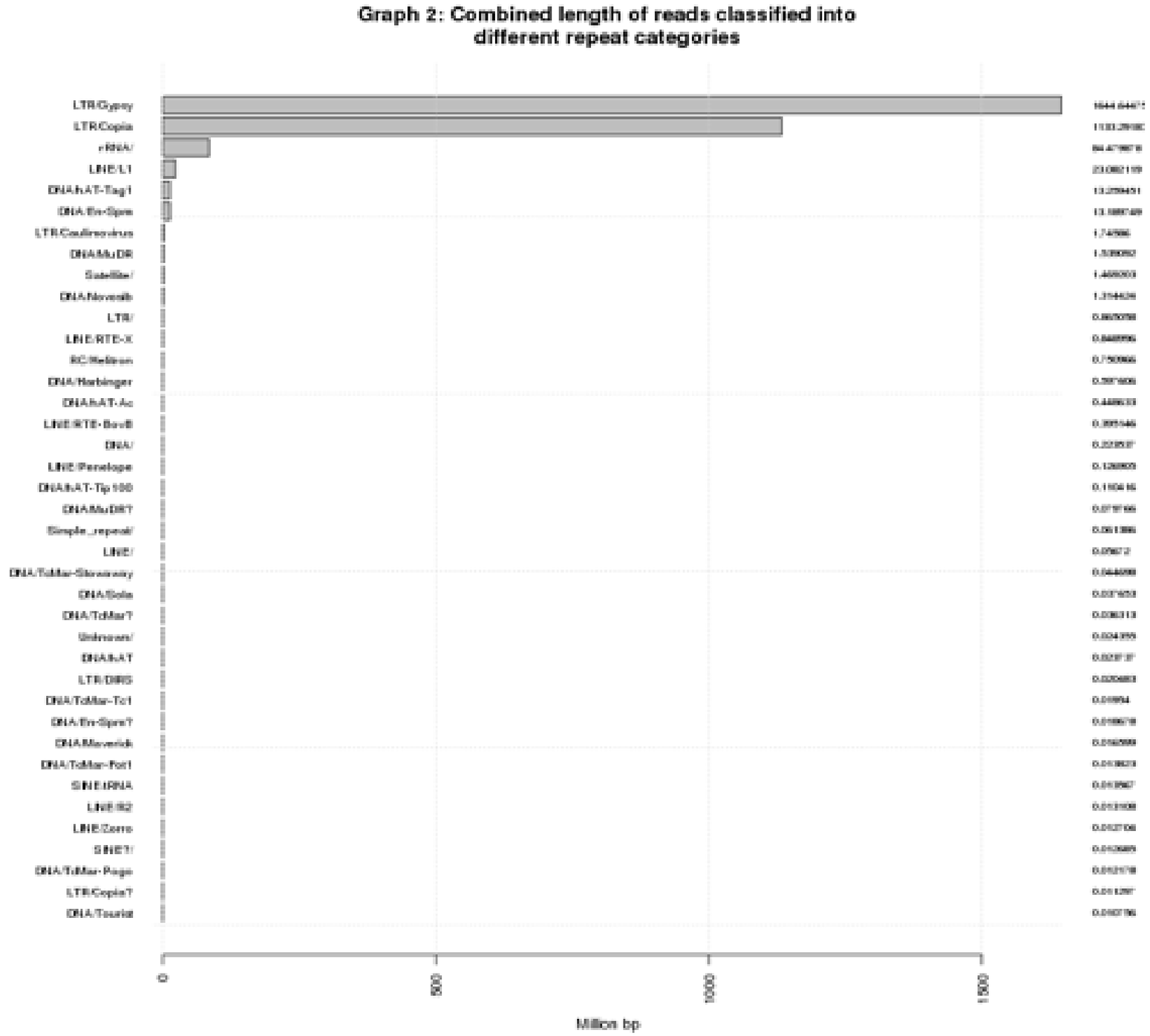
- 5x Illumina sequencing and low-coverage 454 sequencing of five additional conifer species:
  - *Pinus sylvestris* (20x)
  - *Abies siberica*
  - *Taxus baccata*
  - *Juniper communis*
  - *Gnetum*
- Main objective is to analyze major repat content in the different genomes
- *Pinus sylvestris* sequencing will be extended in EU FP7 project “ProCoGen” starting early 2012



The Spruce Genome Project

Analyzing the repetitive part of the genome

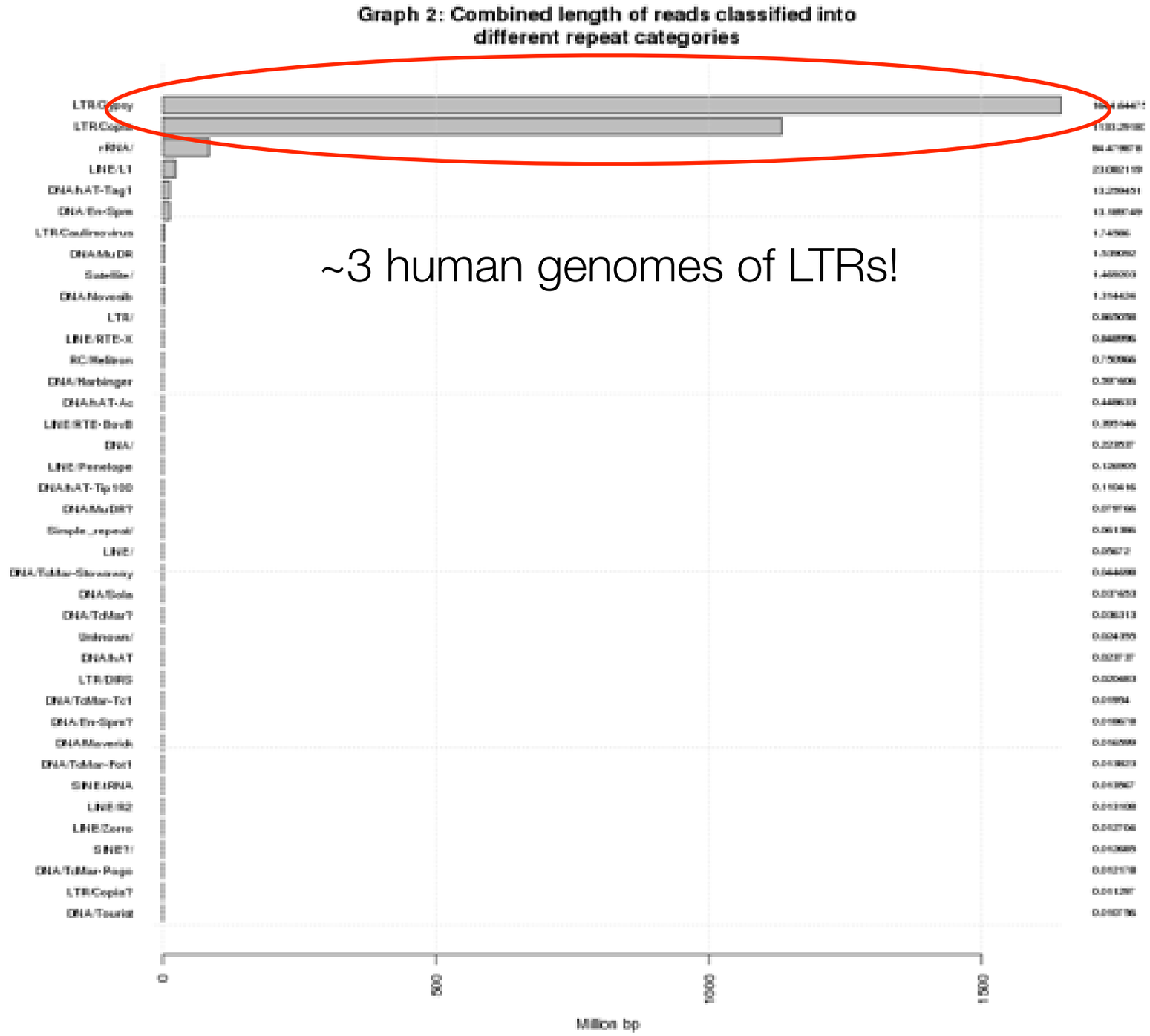
- Using low coverage 454 data to identify and classify repeats
- 1838 repeats identified, 1404 of these can be assigned to known classes
- These repeats mask 69% of the 454 read set



The Spruce Genome Project

Analyzing the repetitive part of the genome

- Using low coverage 454 data to identify and classify repeats
- 1838 repeats identified, 1404 of these can be assigned to known classes
- These repeats mask 69% of the 454 read set



More information on repeats:

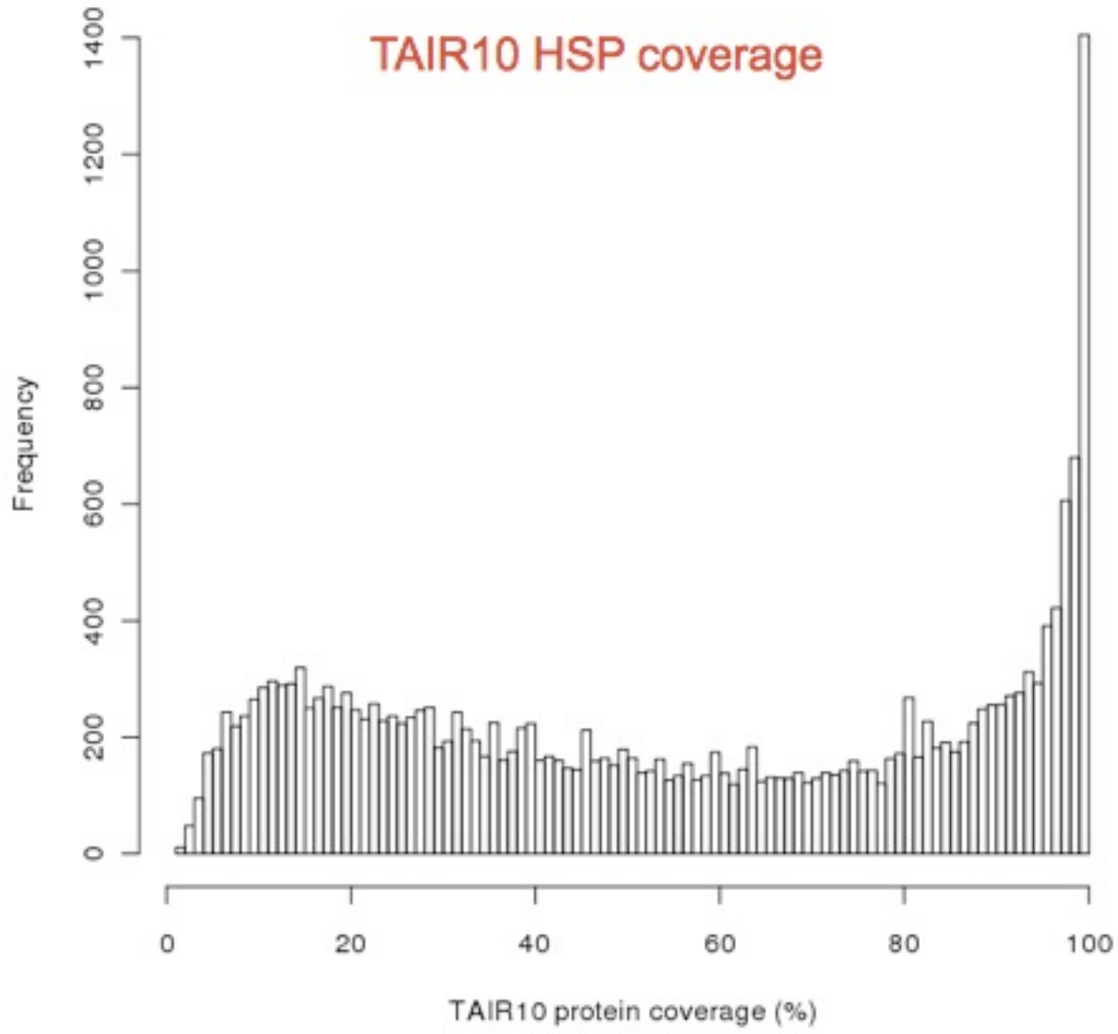
- *Sunday Jan 15, 8:30 am:* The repetitive DNA of conifers - lessons from the Norway spruce (*Picea abies*) genome and resequencing of 5 other conifers - *Anna L Wetterbom*

Transcriptome analyses

- **Phase 1 - gene calling and discovery**
  - 24 tissue / seasonal samples from Z4006
  - Normalised mRNA and total RNA pools
    - polyA selected and random hexamer
    - 454 + Newbler assembly (complete)
- RNA-Seq of all individual samples
  - Illumina strand-specific RNA-Seq

Transcriptome analyses

- Assembly of 454 data
  - 26367 Isogroups / 35992 Isotigs
  - 12,726 in common to Candian white spruce EST set
  - Comparison to TAIR10 shows our 454 EST set is comprehensive and covers as many proteins full-length
  - ~33% of white spruce and 454 Isotigs align within a single shotgun contig and ~66% are well covered but fragmented

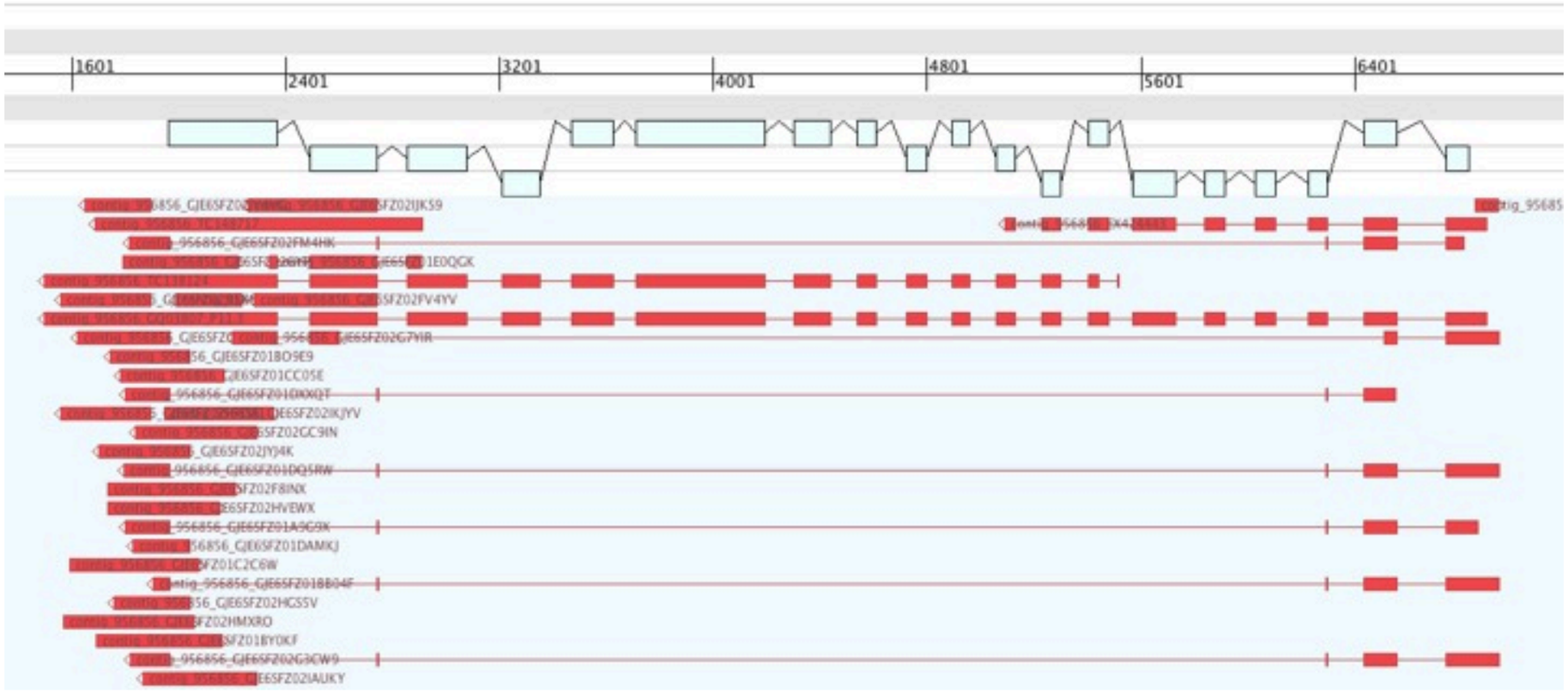


Transcriptome analyses

- Compact genes, with short introns (similar to angiosperm genes)

Gene structure

454 EST

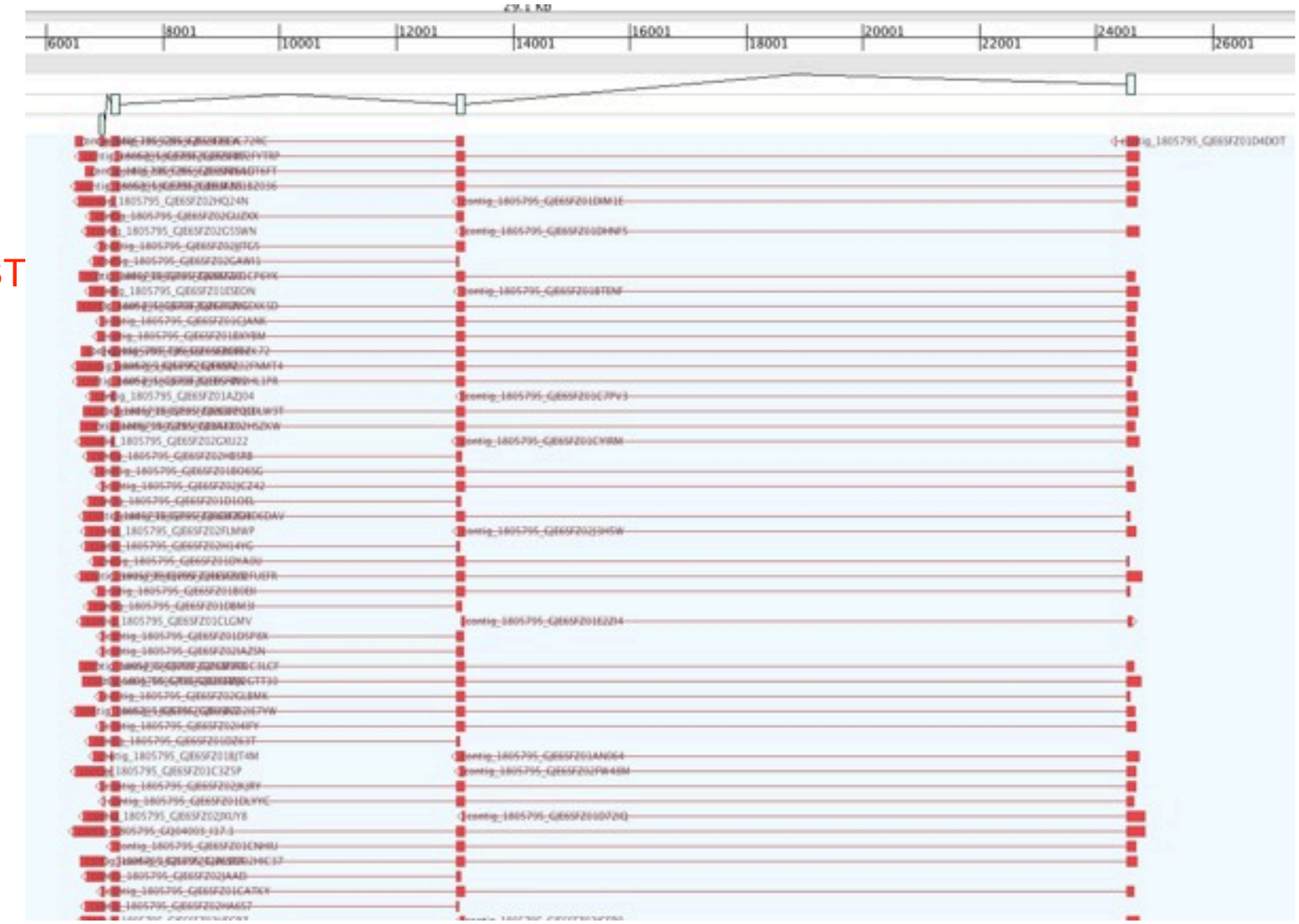


Transcriptome analyses

- Genes with a mixture of long and short introns:

Gene structure

Intron size 123 bp 5,782 bp 11,366 bp



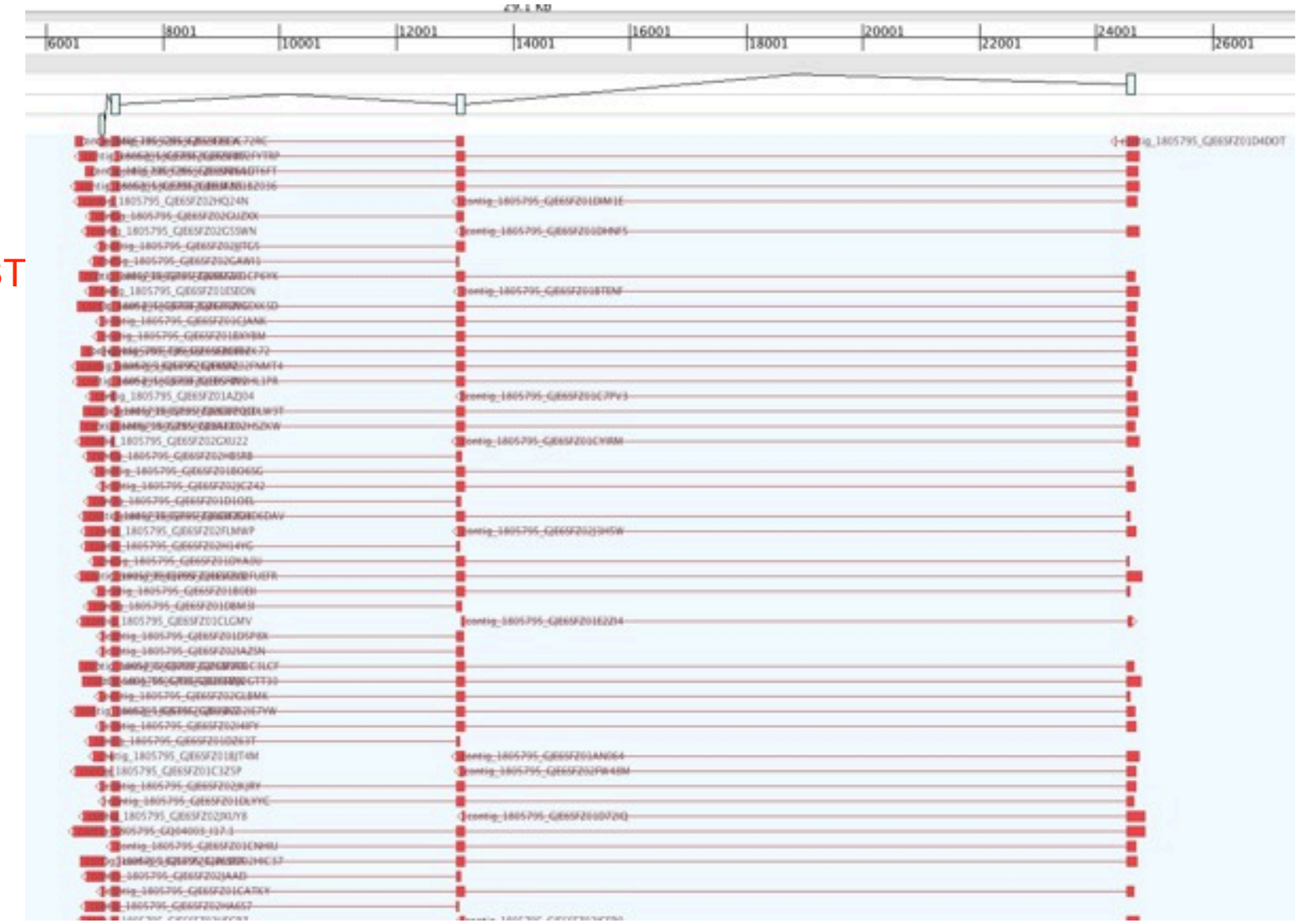
454 EST

Transcriptome analyses

- Genes with a mixture of long and short introns:

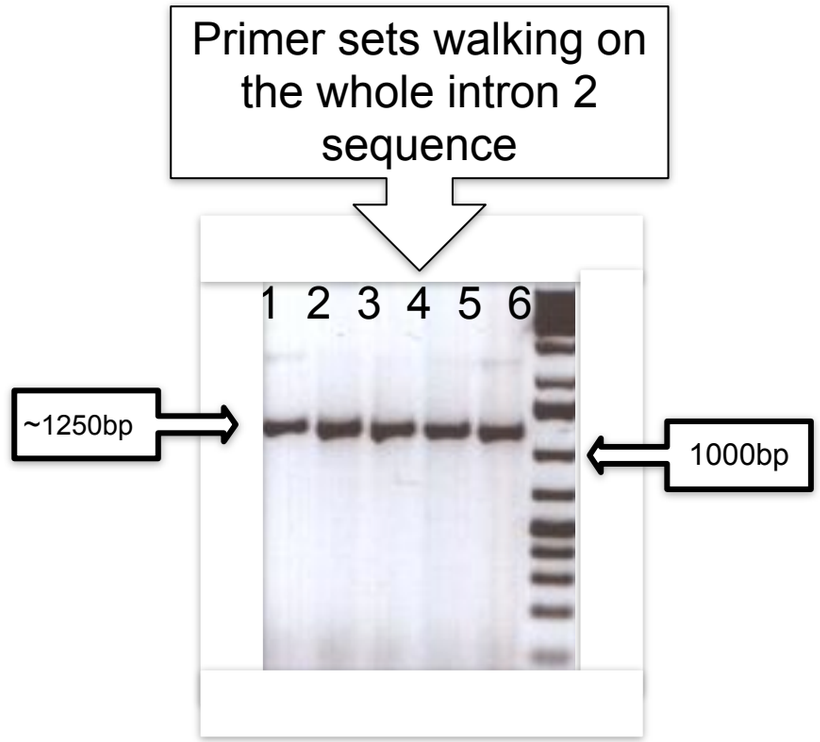
Gene structure

Intron size 123 bp 5,782 bp 11,366 bp



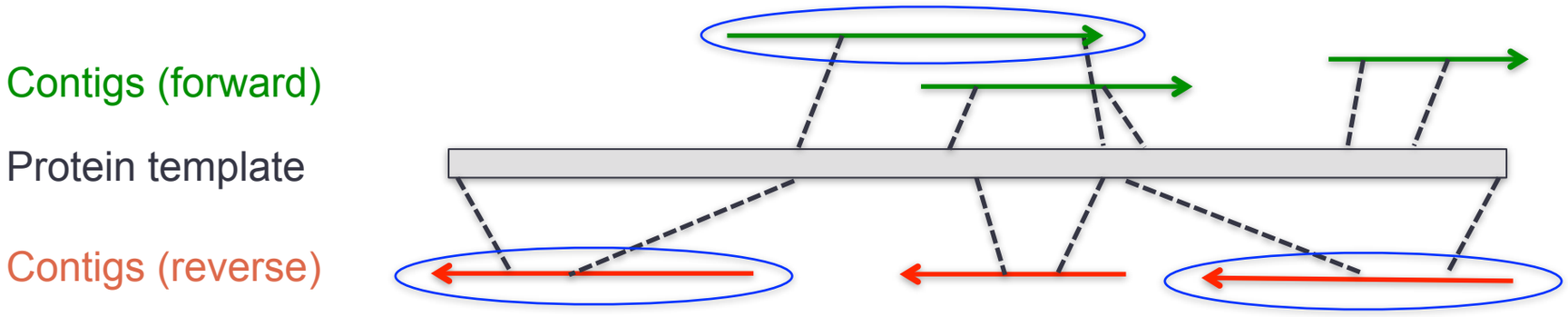
454 EST

Verification:
PCR products (1250 bp) amplified with overlapping primer sets across intron 2 (predicted size 5782bp).

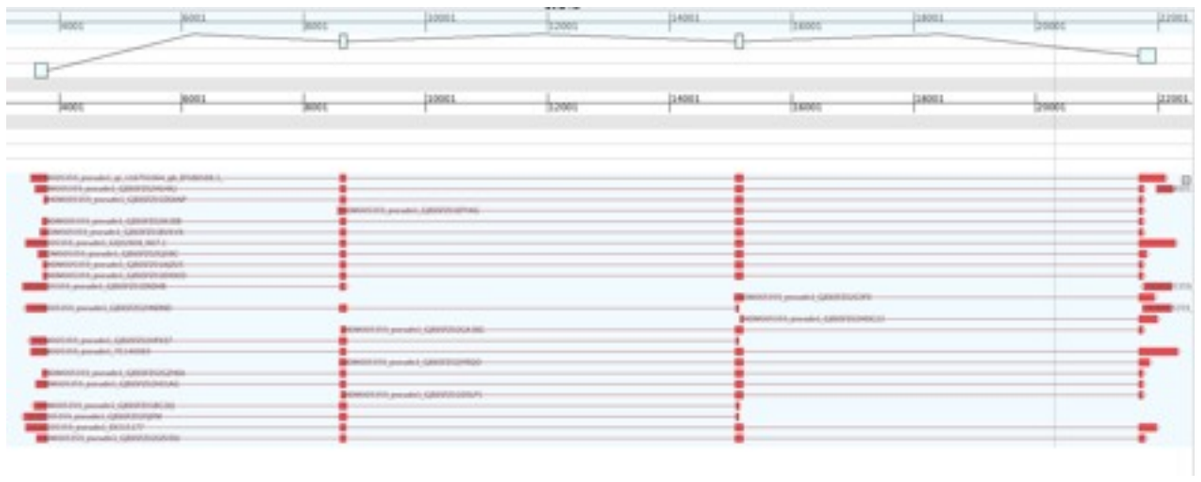


Transcriptome analyses

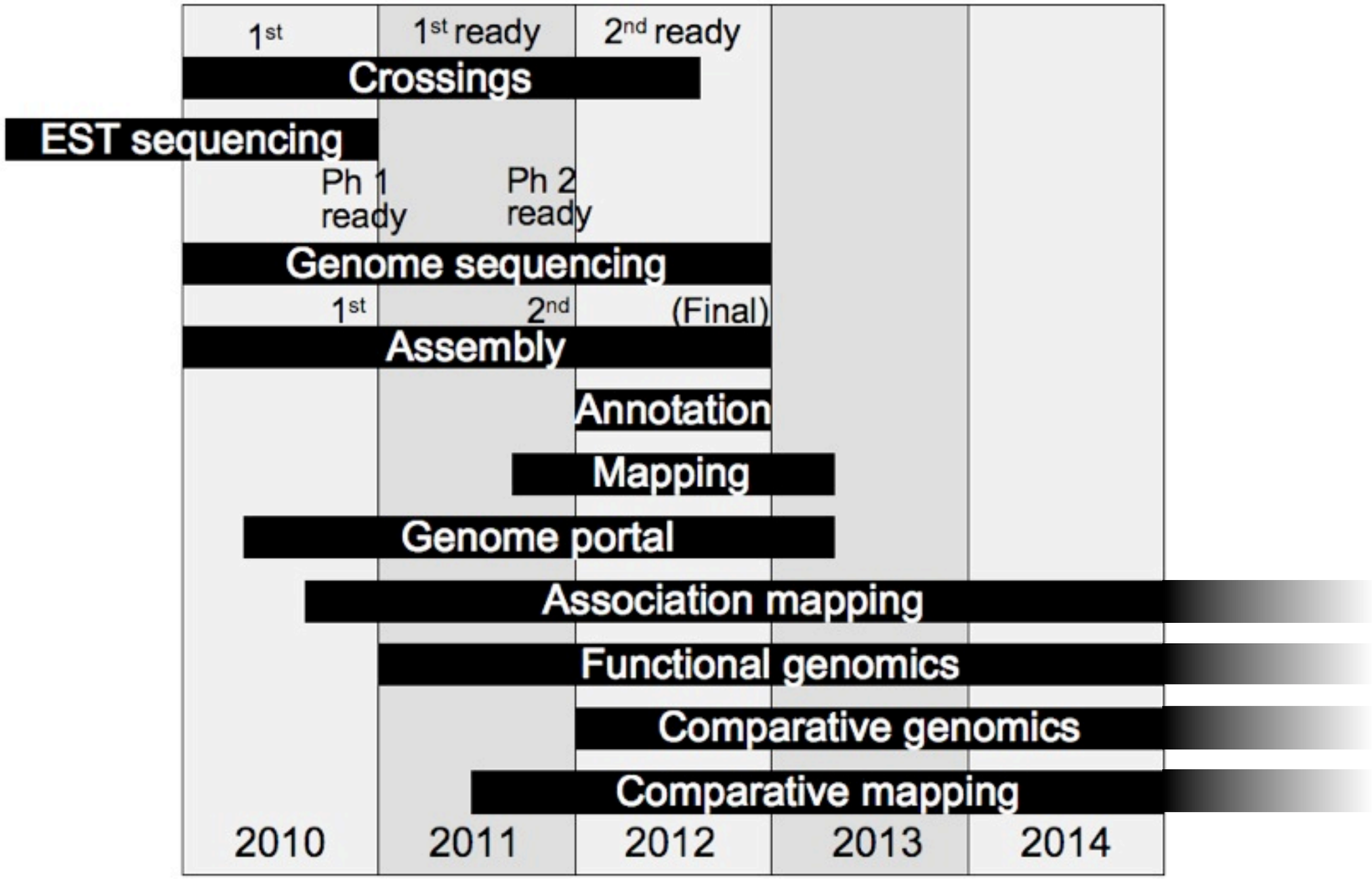
Potential long intron gene scattering in three contigs:



Joining three contigs into one (pseudo)scaffold
then 454 ESTs can properly align to the scaffold



Downstream proof-of-principle projects



Next phase of transcriptome analyses

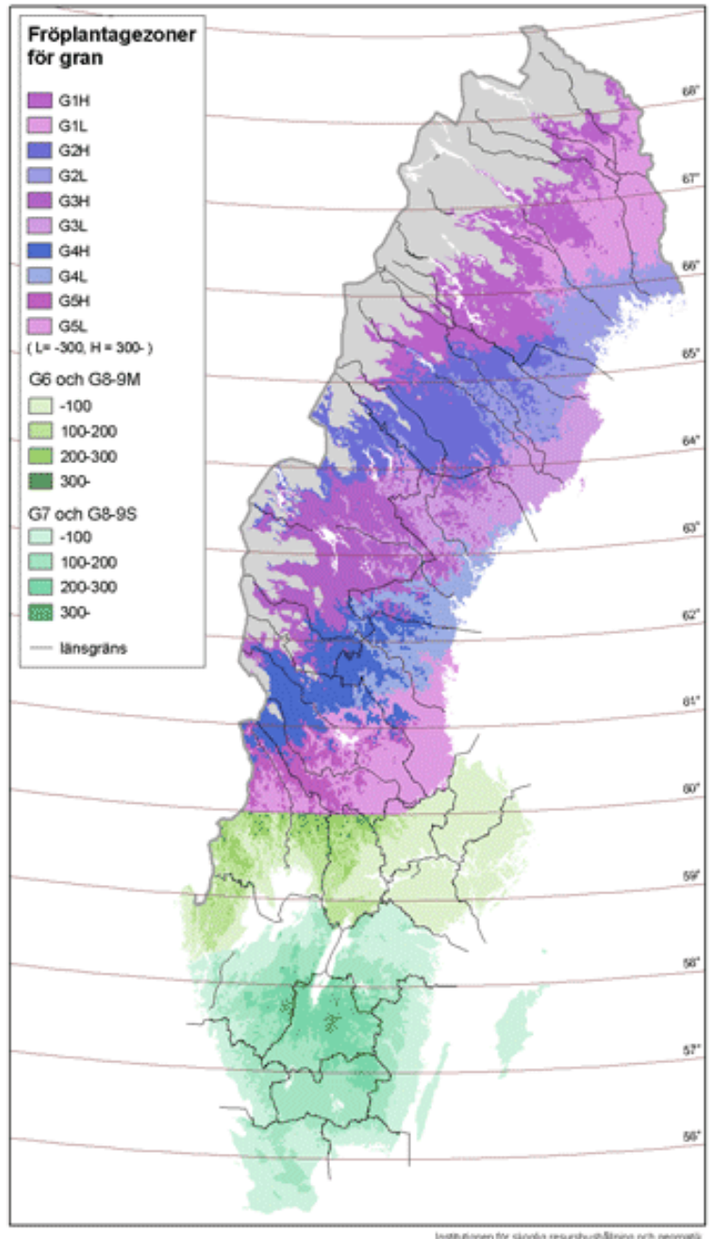
- **Phase 2 - biological discovery**
  - 18 projects, > 800 samples
  - Emphasis on wood development
  - Projects led by a UPSC PI
  - All data publicly available at http://congenie.org
  - Combined systems biology analysis of all data

Related talk:

Sunday Jan 15, 2:40 pm: Large-scale RNA-seq transcriptomics studies Exploring wood development and natural variation in aspen (*P. tremula*): Projects and Resource Development - *Nathaniel Street*

Downstream proof-of-principle projects

- Association mapping
  - Open-pollinated families from SkogForsk's breeding trials, two replicated populations of roughly equal size (1200 families per replicate)
- Functional genomics
  - Spruce somatic embryogenesis and transformation.
    - Analysis of selected transcription factors (in collaboration with STT)
- Comparative genomics
  - Spruce – Pine comparative mapping
  - Spruce – *Populus* - *Arabidopsis* comparisons



Making the data available:

ConGenIE - the Conifer Genome Integrative Explorer

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

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

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

CLCbio

Lucigen

Knut och Alice Wallenbergs Stiftelse
Umeå Plant Science Centre a centre of excellence
SciLifeLab
IGA
VIB
Umeå Universitet
SLU
KTH
Stockholm University
Karolinska Institutet
Uppsala Universitet
C·H·O·R·I Children's Hospital Oakland Research Institute
SKOGFORSK