# 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





### 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 treebased 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.



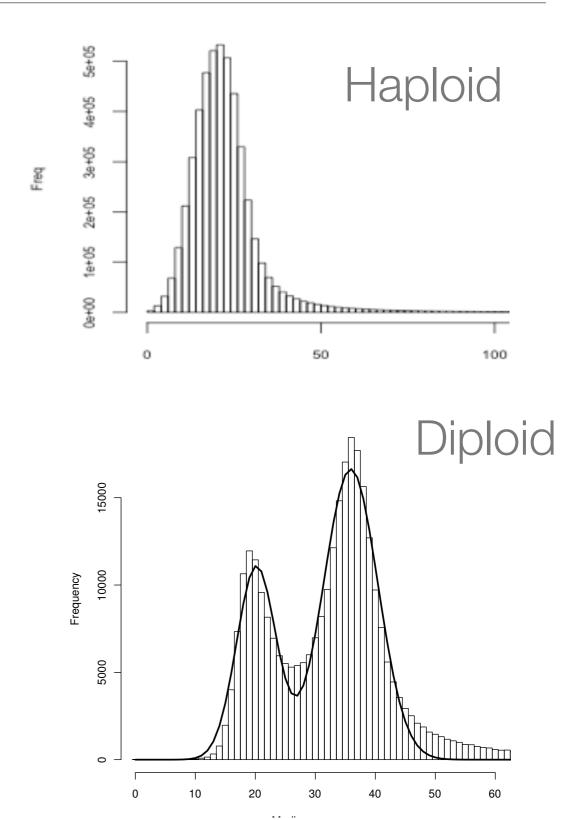


# Sequencing status

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 ends Fosmid pools	20X span cov 500 000 40 kb fosmids in pools of 1000
	500 000 40 kb fosmids in pools of 1000

### 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			



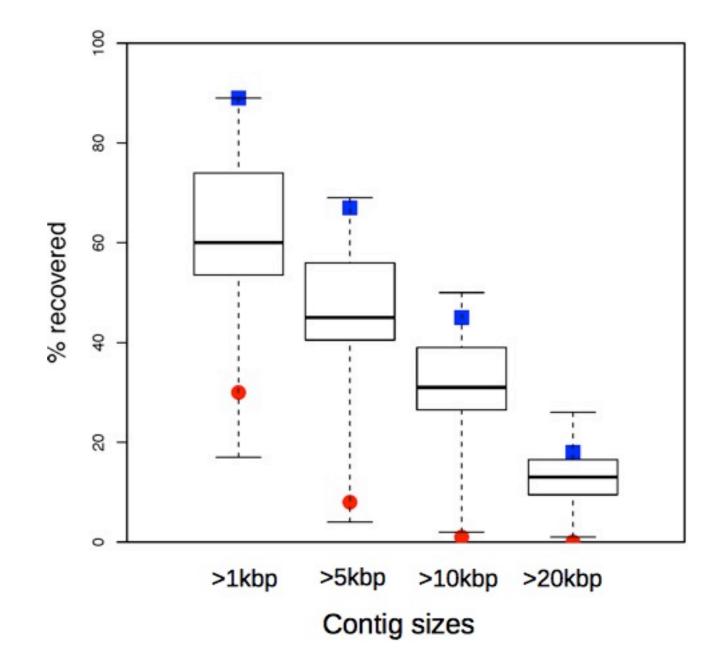
# Fosmid pool strategy

#### **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

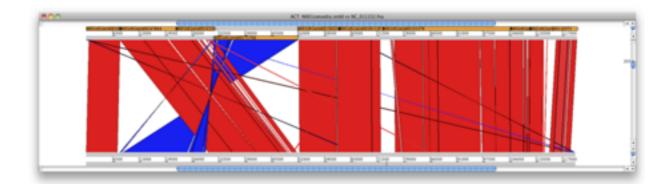
#### Sequencing the *Populus tremula* genome:

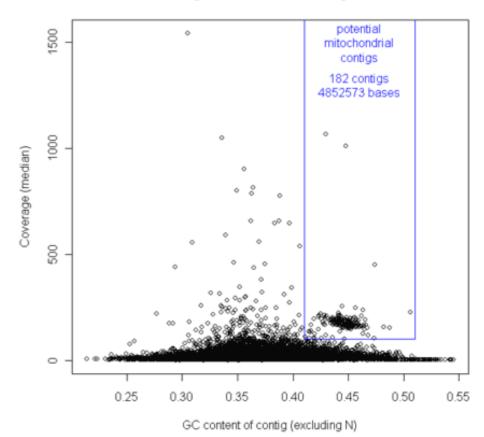
• Sunday Jan 15, 11:30 am: The genome of aspen (Populus tremula) - the most complex genome sequenced to date? - Stefan Jansson

# 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?



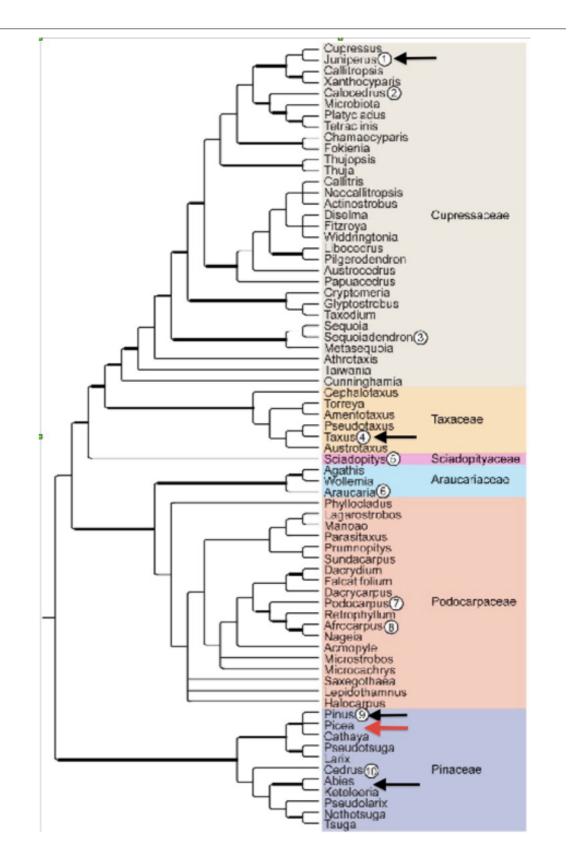




#### Coverage vs. GC for all contigs > 5kb

#### 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





### 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

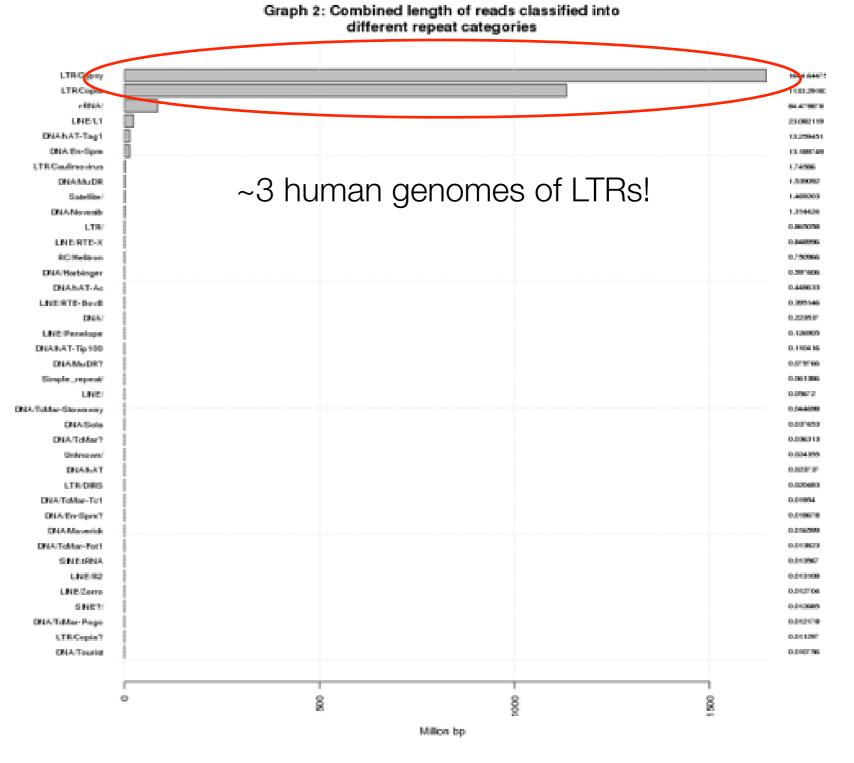
he Spruce Genome Project

R Gypry			
R Copia			
r RNA/			
LN611			
AT-Tag1			
LNEL1			
inovirus			
AMUDR			
Sudellike/			
Novesile			
LTR/			
ERTE-X			
Nelleon			
orbinger			
hAT-Ac			
E-Bevil			
DRAV			
enelope			
-Tip 100			
MuDR7			
,repeak'			
LNE/			
unanany .			
NA Sola			
TdMar?			
itnows/			
RABAT			
TR DIRS			
Mar-Tc1			
n-Spm?			
Maverick			
Ase-Pot1			
VEXPNA.			
LNER2			
(E/Zerre			
SNET			
ar-Pogo			
Copis?			
AlTourist			
		1	i
0	8	000	1800

Graph 2: Combined length of reads classified into different repeat categories

### 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



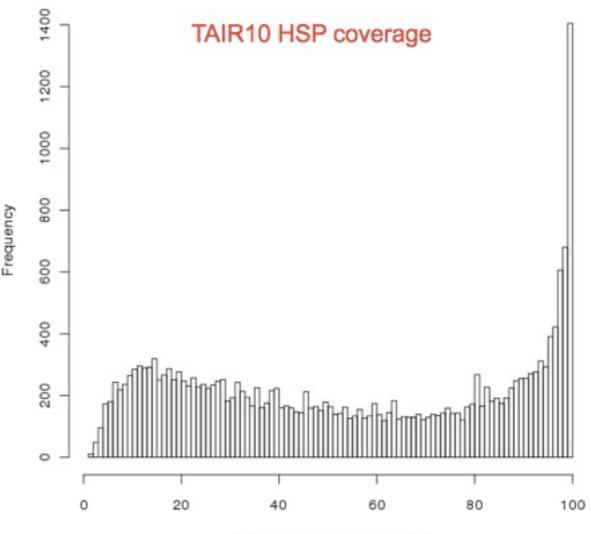
• Sunday Jan 15, 8:30 am: The repetitive DNA of conifers - lessons from the Norway spruce (*Picea abies*) genome and resequencing of 5 other conifers - Anna L Wetterbom



- 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



- 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



TAIR10 protein coverage (%)



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



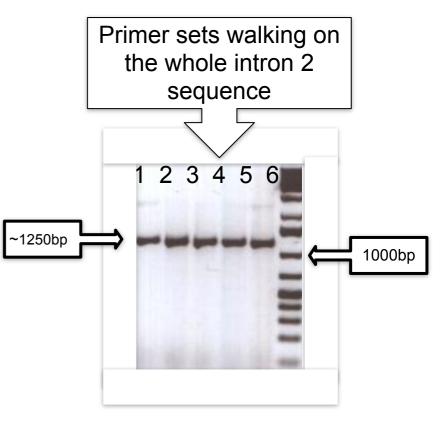
- Genes with a mixture of long and short introns:
- Gene structure

itron si	<b>Ze</b> 123 bp 5,782 l			р	11,366 bp						
	6001	8001	10001	12001	14001	16001	18001	20001	22001	24001	26001
		n								0	
	-	AND THE SOURCE AND THE	06.70M							der in 18	05795_C2865F201D4D4
	(100 High	Breest Construction	RECEVER								
		1415 200 CR65 2000									
		1805795_C(E69/202HQ			swig_1805795_Q8697201	DIM 16					
		1805795 CREST202									
		1.1805795_GEE5FZ02G		4	centig_1805795_G865F20	obeers.				-	
		1805795_GEK9720 9_1805795_GEK97202									
54 EST		Control Discourse of Figures		-							
		2_1805795_C(065F2110	ISEON		ontig_1805795_G885F204	INTERNE				-	
		ant grant and a statem									
		1805795_C8889780									
										_	
		Banages approved the								-	
		005225000000000			the second concerns	La resulta de					
		9_1805795_CJEE5F201/			1805795_G865F20	00/1743					
		14410/15/0215/021A									
		1805795_GE65F202GX			ontig_1805795_C@65F203	CYAM				-	
		1805705_G885#202#									
		1405795_GR697204 1405795_GR697204									
		- 1805795 CHEST2010									
	-	1000 TRUSPOS ADDR	REPERENT								
		1805795_C(065F202F)			owig_1805795_QE697203	13945W					
		1405795_Q0057202H 8 1805795_Q0057202H									
		1415795_QE65F20	16(0)							- F	
		-1805795_QH65F2010			AND DESCRIPTION OF CARDING						
		1805795_CERSF201CL nig_1805795_CERSF20			contig_1805795_GE65F20	16204					
		nig 1805795 GE65F25									
THE R OTHER PROPERTY AND A	PRODUCILOF										
		out to character									
		Hig_1806705_C(EKSF20 Melliphic)C(EKSF20)C(EKSF2									
		tig 1805705 Catabilit									
		1805795 GRISTING		ī							
Generation (1805795, GRESF2018)74M     Generation (1805795, GRESF201C125P			ontig_1805795_C(665720)								
		untig_1805795_QE65F20	2110-4550								
	Brig_1805795_GE65F202808Y     Brig_1805795_GE65F20100YYC								-		
	-	1805795_C(665F202)X	178	4	contig_1805795_Q865720	05720					
		805795_CQ04003_137									
		ontig_1805795_CJE65 13666420542689526346									
		-1801795-CJE6572029									
		aig 1805705 CHEST21		-							
	1000	1805795 CAMIFICAN	A657								

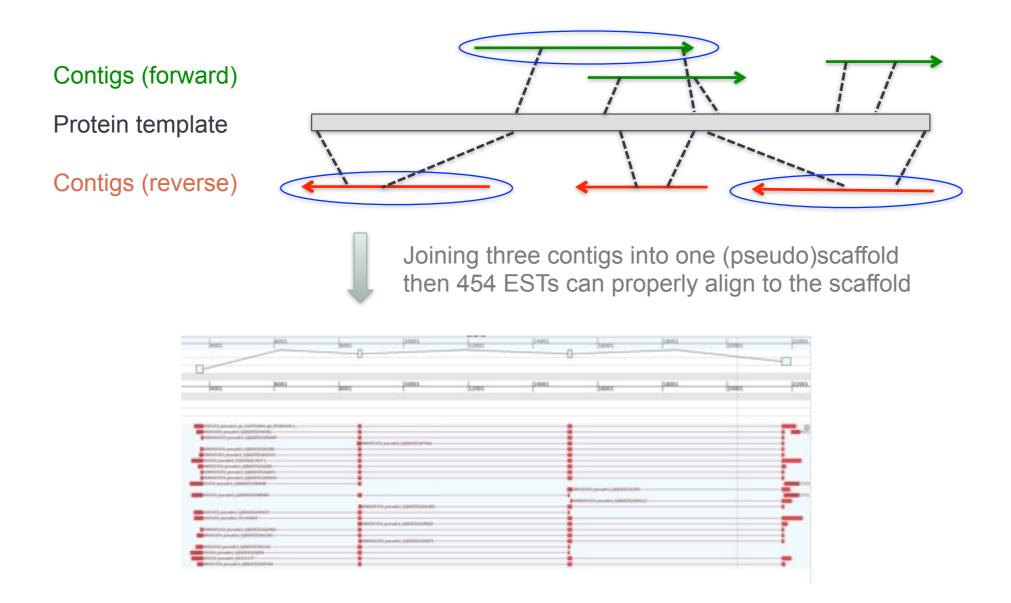
- Genes with a mixture of long and short introns:
- Gene structure

ntron siz	ze 12	3 bp	5,782 k	р		1	1,366 bp				
	6001	8001	10001	12001	14001	16001	18001	20001	22001	24001	26001
					1						
	N										
	ď									0200	
		AND TRACESSORY AND A								4+ <b>2</b> 10_180	05795_CJEESF203D40
		International Content of Longian Content									
		ANGELEVENE COMM		-						-	
		1805795_C(E69/202HQ 1805795_C(E69/202HQ			swiig_1805795_Q865F201	CAW IE					
		1805795_C865F202G			untig 1805705_GE65F20	UDHINES					
		149 1805795 GRESF20		-						100	
EA LOT		2_1405795_GB657202								1.1	
54 EST		hereity in syner speed			and the second	ATTAK					
		2_1805795_C(E65F2016 4465_2)1005F2016			ontig_1805795_C(845720)	IN TONO					
		1805795 Calestra									
		ig_1805795_Q869720		-						-	
		4405700-196-60166A									
		846-871-926-992-6269488									
		g_1805795_CJEE5F2014			ontig_1805795_G665F20	167Pv3					
		1446/ 18/15/295/ Epoca	REGIDLW97	-		1.000				-	
		MANUTACIAN CRIAT		-		2010 C					
		1805795_GE65F202CX 1805795_GE65F202CX			ontig_1805795_C(865920)	CYRM					
		nig_1605705_C(665720									
		- 1806395_GEE972010									
		1805795 CJD65F202FL			ania lanting concerns	a last la					
		1805795 G86657202H			untig_1805795_QE697203	19-04					
		a 1805795 QE697201		-							
		10167-2040/005-20464		-						-	
		ig_1805795_QE65F20								-	
		1805795_G865F2010 1805795_G865F203CU			ontig 1805795_QE69520	HERE .					
		nig_1805795_G865F20		-	and a second second						
	44	nig_1805795_QBISF20	Q1425N								
		pany property cares									
		oang 160 Citer Citer 20									
		MARCOL REPORT GROUP									
	44	tig_1805795_C(685720	QHEFY.							-	
		1805795_Q86572010								_	
		tig_1805795_CJ065F201 1805795_CJ065F201C32			untig_1805795_C2005F201 untig_1805795_C2065F201						
		1805795 C2055720			and many of the set	CT 40 4 200					
		Hig 1805795 QESSER								-	
		1805795_C#65F202p0		4	contig_1805795_G(665F20	06250				-	
		905795_CQ04003_107									
		Intig_1805795_C(8657 [1004203.42679526]003								100	
		1805795 CJE65F20254									
	44	1801705_C(E65720	HCATER	-							
		1805795 Category 20200	4657								

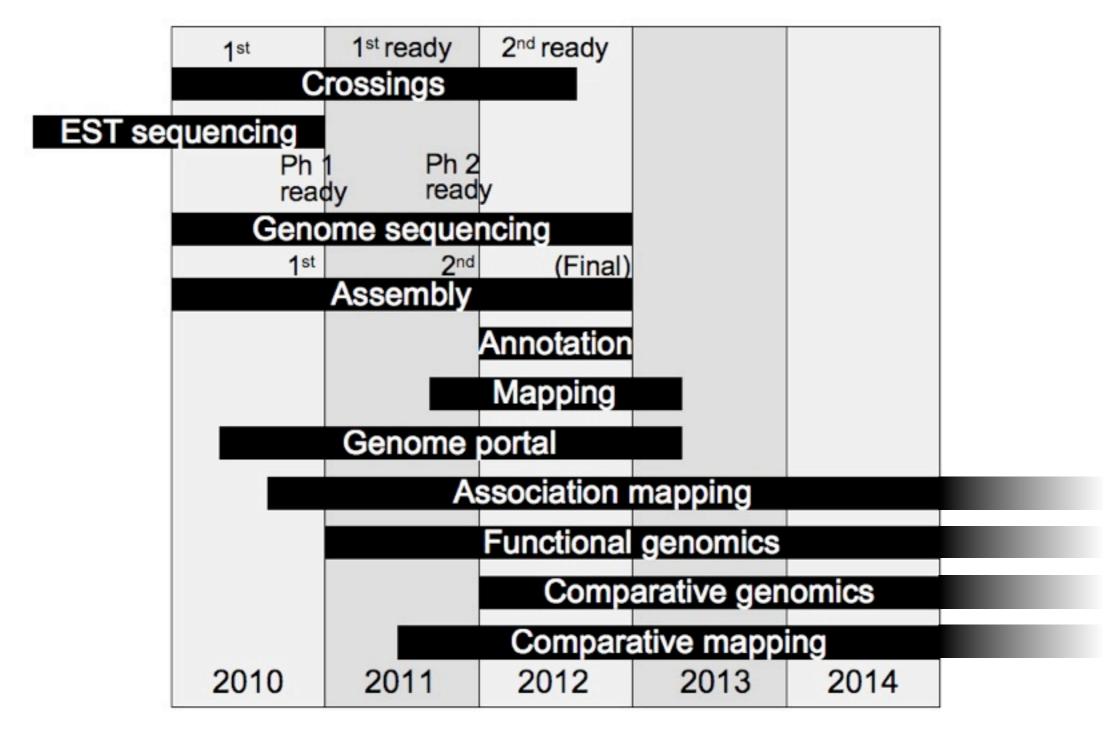
Verification: PCR products (1250 bp) amplified with overlapping primer sets across intron 2 (predicted size 5782bp).



Potential long intron gene scattering in three contigs:



### 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 <u>http://congenie.org</u>
  - Combined systems biology analysis of all data

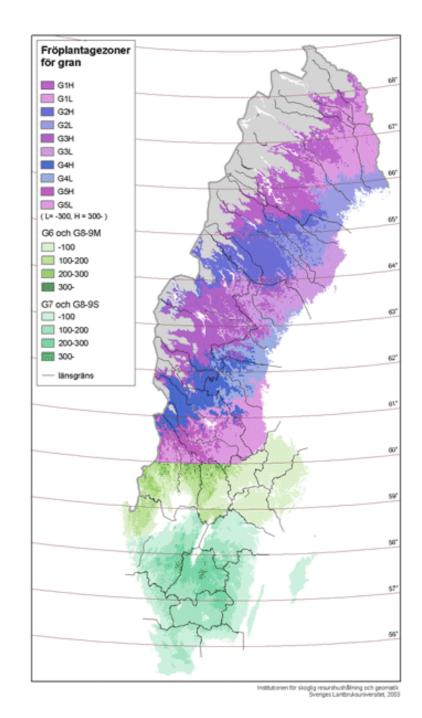
#### **Related talk:**

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

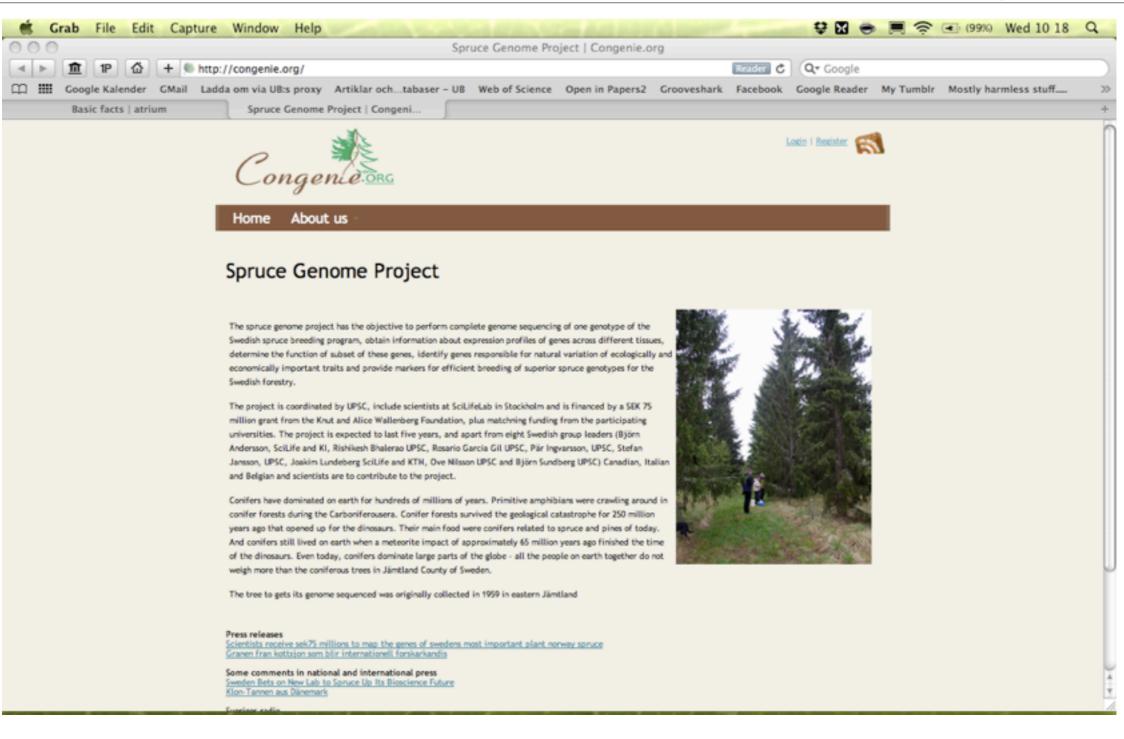
- Association mapping
  - Open-pollinated families from SkogForsk's breeding trials, two replicated populations of roughly equal size (1200 families per replicate)
- Functional genomics
  - Spruce somatic embryogenesis and transformation.
  - Analysis of selected transcription factors (in collaboration 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

#### SAB

Kerstin Lindblad-Toh John MacKay Outi Savolainen Detlef Weigel



VIB Gent Yves Van de Peer Yao-Cheng Lin

**Skogforsk** Bengt Andersson Bo Karlsson **IGA Udine** Michele Morgante Francesco Vezzi Ricardo Vicedomini Andrea Zuccolo

SNIC Supercomputers Uppmax/PDC/NSC/HPC2N **CHORI Oakland** Pieter de Jong Maxim Koriabine

SNISS national infrastructure

Karolinska

Institutet

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













 $C \cdot H \cdot O \cdot R \cdot I$ 



SKOGFORSK