The Spruce Genome Project

SciLifeLab

Sequencing and assembly of the largest and most complex genome to date

The Norway spruce (*Picea abies*)

Björn Nystedt, SciLifeLab, Sweden
The Spruce Genome Project

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

In particular for this talk..

SciLifeLab

Data analysis
Douglas Scofield
AndreyAlexeyenko
Anna Wetterbom
Ellen Sherwood
Nat Street
Yao-Cheng Lin

Scaffolding (BESST)
KristofferSahlin

Fosmids
Pieter dJ (CHORI)
Lucigen

Sequencing
SciLifeLab Genomics Platform
PacBio

Assembly software development
CLCbio

Computing and storage
Uppmax (30 TB disc and counting..)
SciLifeLab (2TB RAM)

SciLifeLab

5 HiSeq
5 SOLiD
3 454
2 IonTorrent
1 MiSeq

Genome size

Arabidopsis
(0.12 Gbp)

Populus
(0.45 Gbp)

Humans
(3 Gbp)

The Spruce Genome Project

Genome size

Arabidopsis
(0.12 Gbp)

Populus
(0.45 Gbp)

Humans
(3 Gbp)

Spruce
(20 Gbp)

The Spruce Genome Project

Spruce as a new model species

Economicallyimportanttree

- Tools for breeding for tree productivity, quality, health
- Tools for cellulose and woodfibre modification (new materials)
- Tools for tree-basedbiorefineries

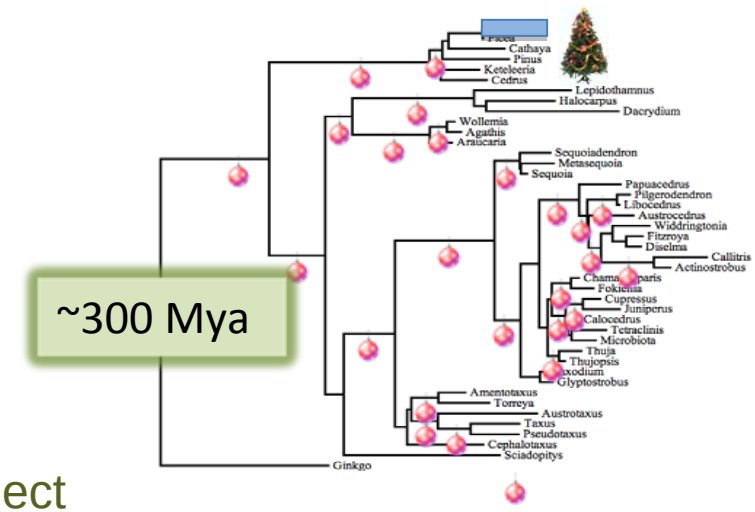
Science

Science of conifers

- Evolution: The last major plant group without a sequenced genome

•Ecology: Dominant members of boreal forests

•Biology: Uniquebiological features

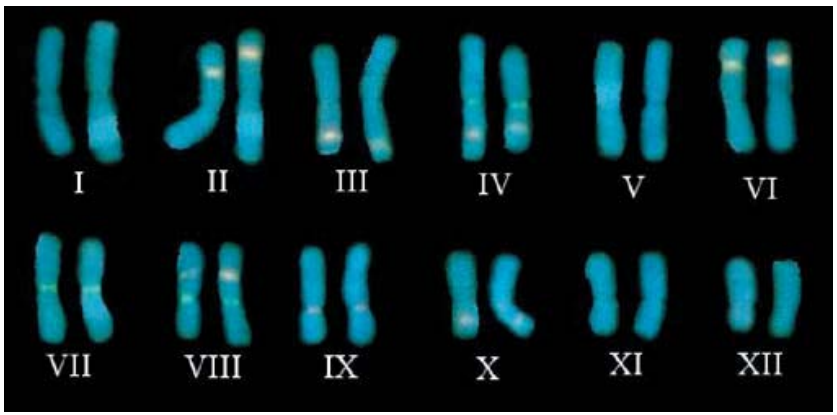


The Spruce genome

SciLifeLab

Challenges

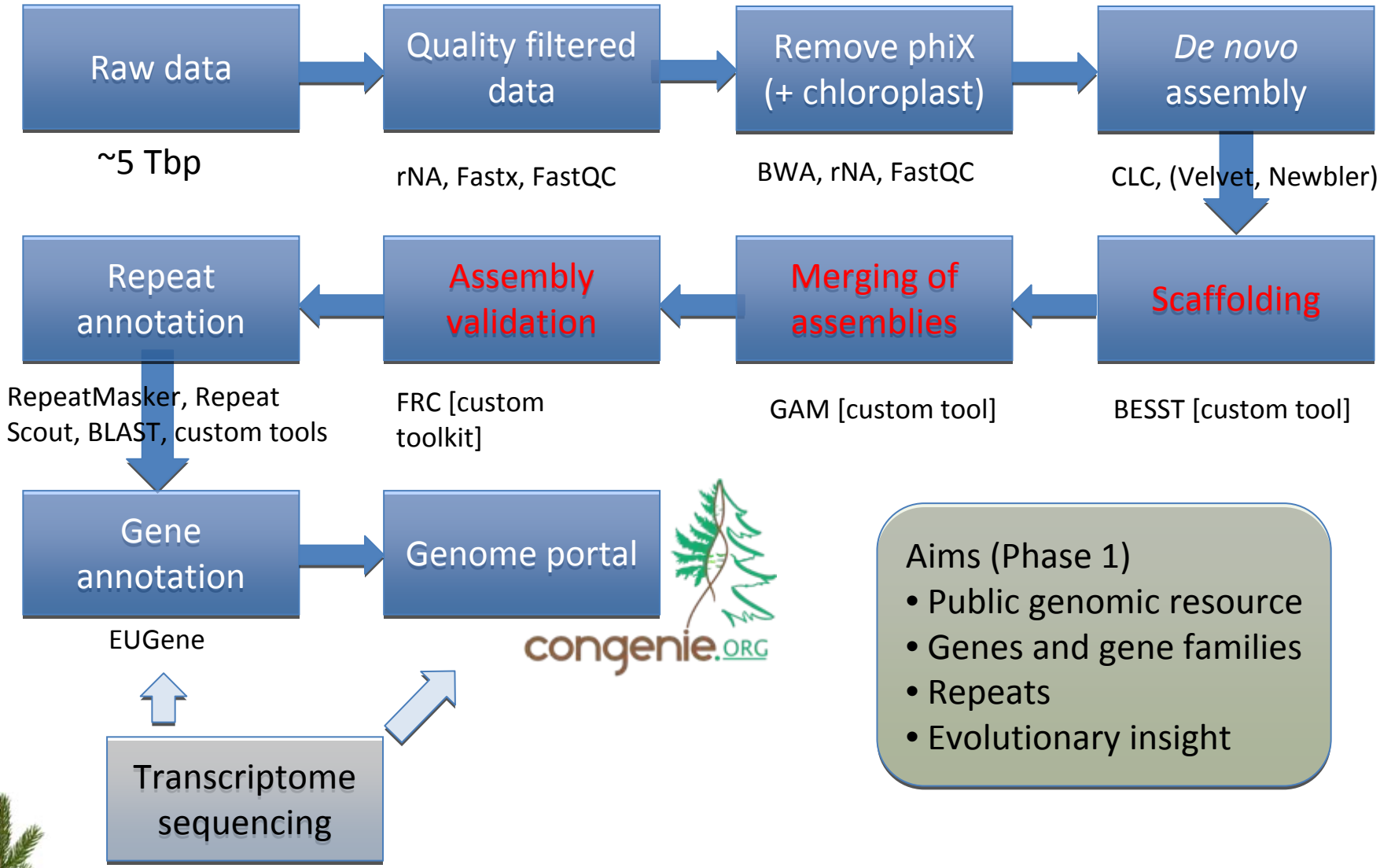
- 19.6Gbpgenome (40% GC)
- 12 x 2 evenly sized chromosomes
  (Chromosome sorting very difficult)
- 75% of the genome consists of transposable elements
- 3% consists of genes and pseudo-genes
  (Large gene families and many pseudo-genes)



Vischi et al (2003)

What do we do with all the data?

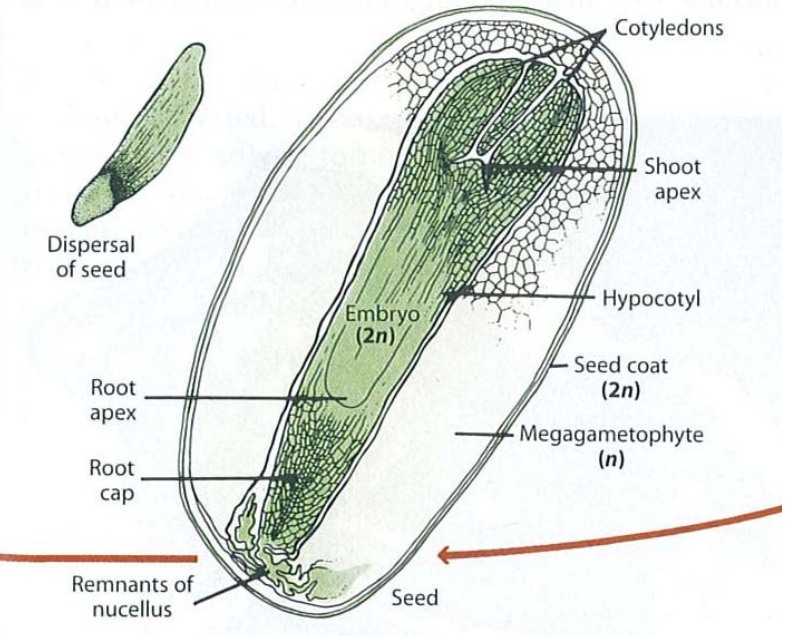
SciLifeLab



Aims (Phase 1)

- Public genomic resource
- Genes and gene families
- Repeats
- Evolutionary insight

The haploid megagametophyte



The megagametophyte (seed nutrient tissue) is haploid from the mother.

~600 ng DNA

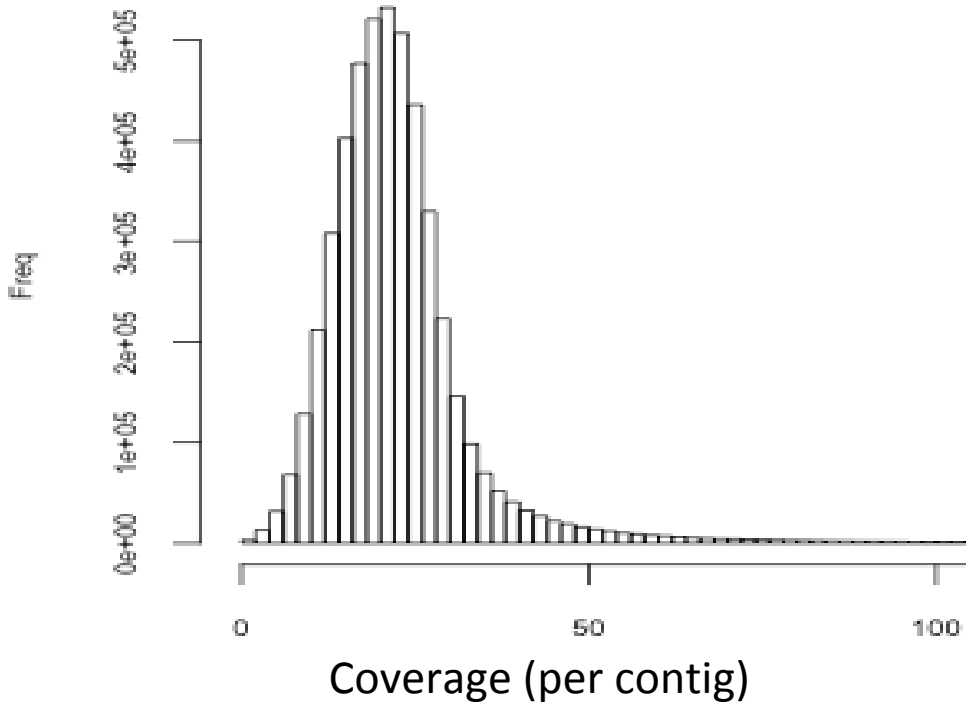
WGS (haploid)	**Status**
PE (150bp, 300bp, 650bp)	20X

WGS (haploid)

Assembly stats, 20X haploid

- 30 % in contigs>1 kbp
- 8 % in contigs>5 kbp
- 1% in contigs > 10 kbp

- NG50: 204 bp



BUT…

Low amount of input DNA leads to library depletion:

=> True coverage is only 10X (50% PCR redundancy)

WGS (diploid)

Needles (normal diploid tissue)
Amount of DNA is not a limitation

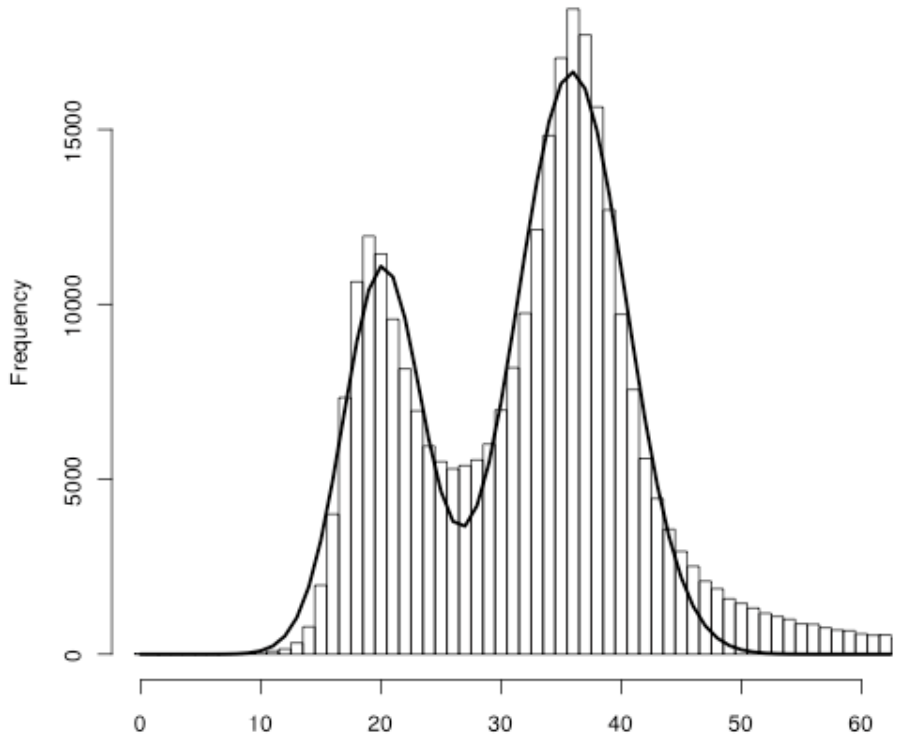
WGS (diploid)	**Status**
454 (SE)	1.5X
PE (150bp, 300bp, 650bp)	50X

~10 billion reads
CLCbio: 5 days (800 GB RAM)

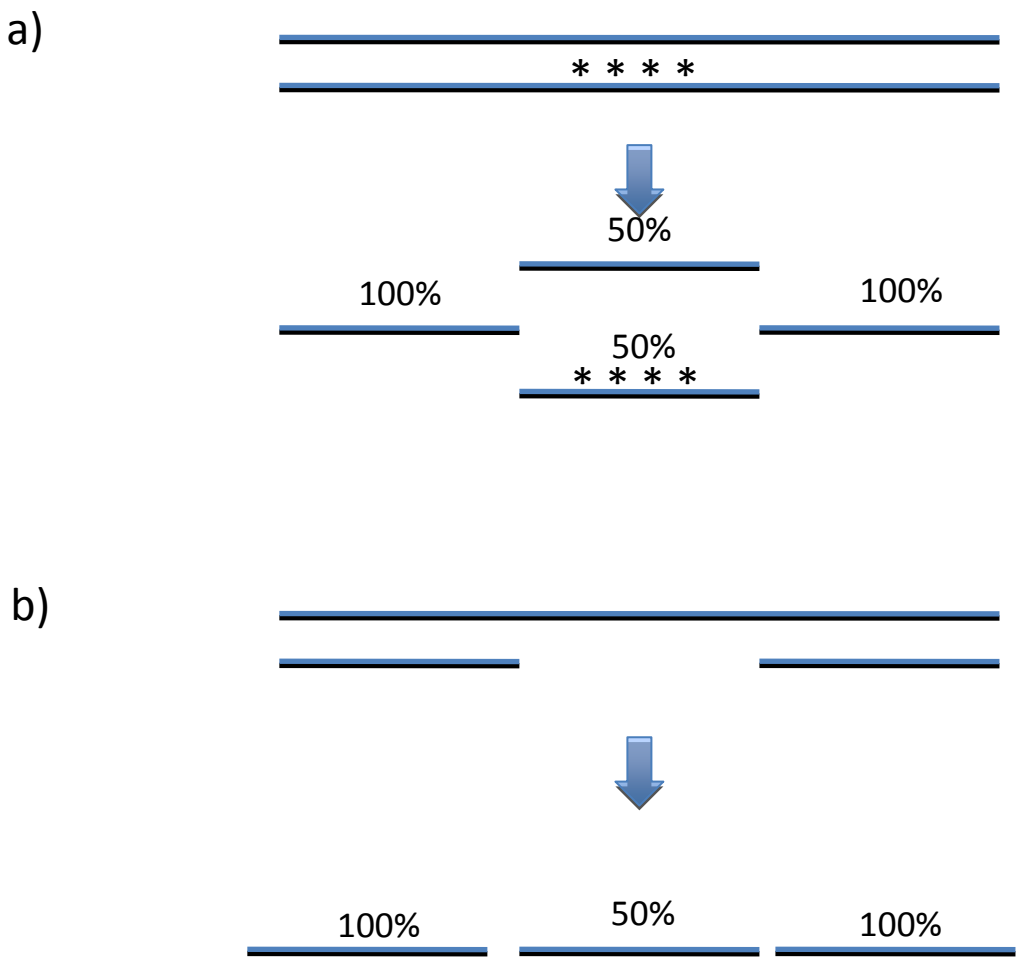
Assembly stats, 50X diploid

- 44 % (30 %) in contigs >1 kbp
- 12 % (8 %) in contigs >5 kbp
- 3 % (1 %) in contigs > 10 kbp

- NG50: 757 bp(204bp)



Why 50% expected coverage?

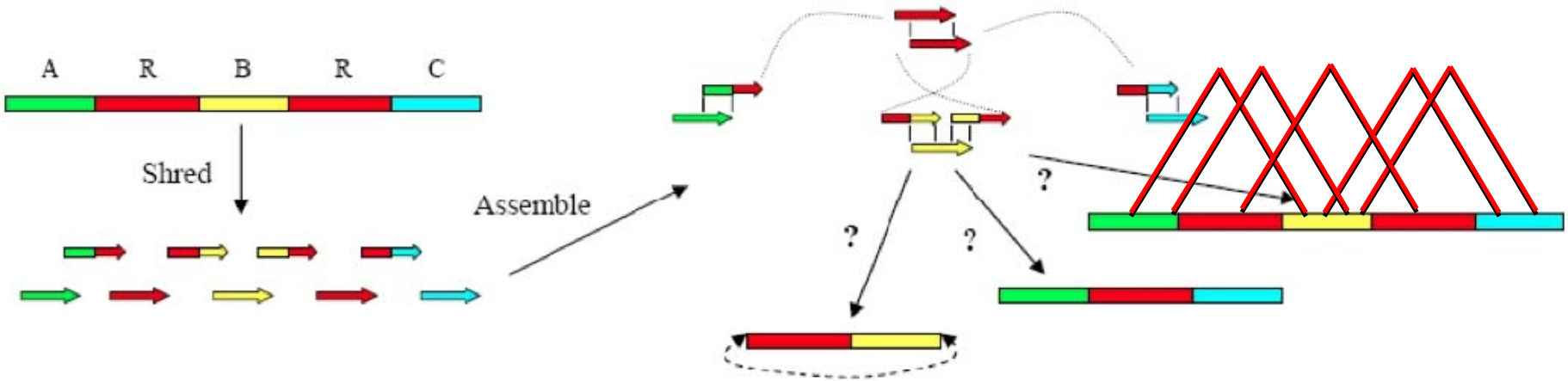


Scaffolding with paired reads

SciLifeLab

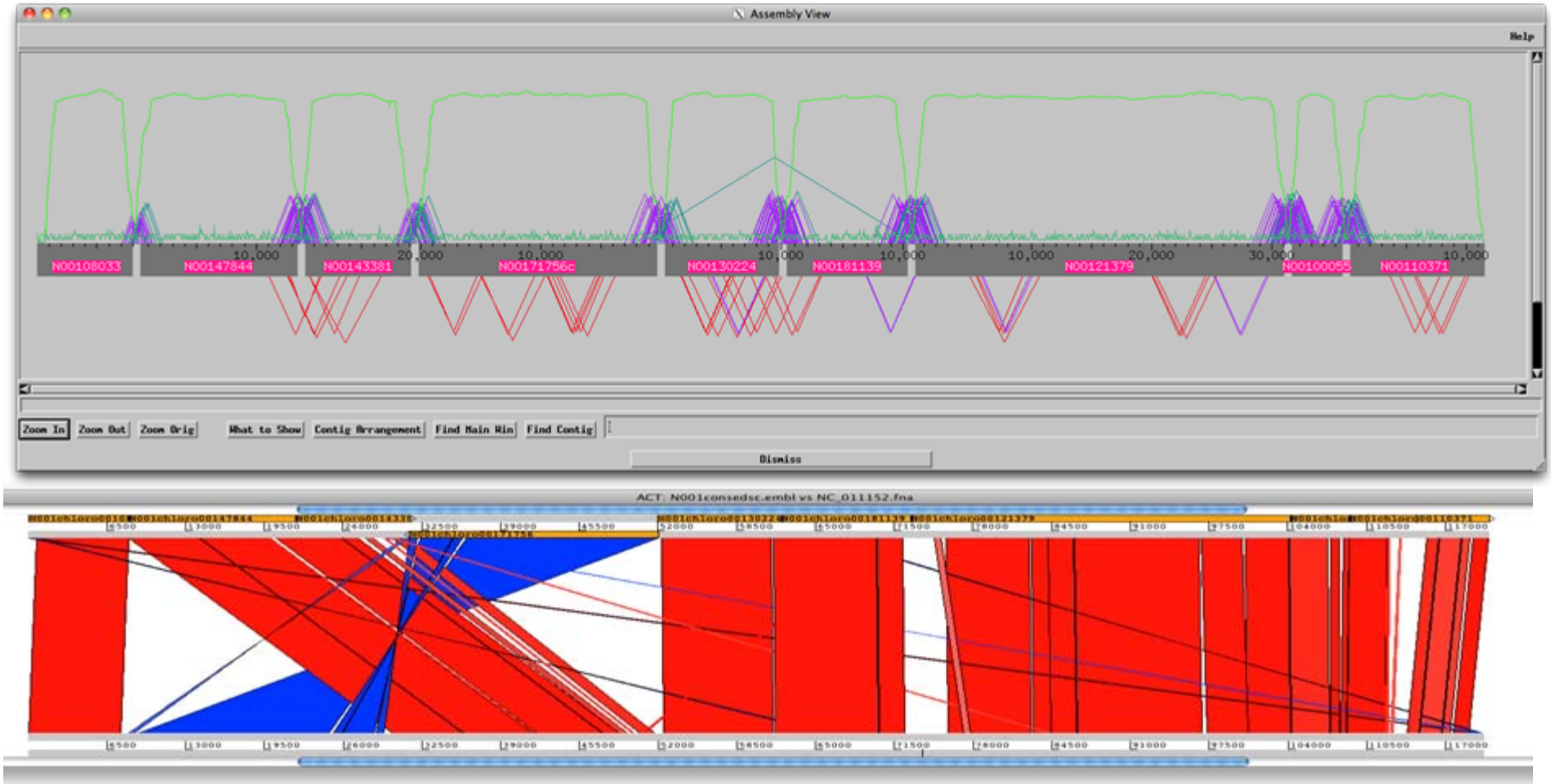
Reasons for contig breaks:

- Repeats
- Local lack of coverage
- Polymorphisms (only diploid)



In-house scaffolder, BESST
“quality over quantity”

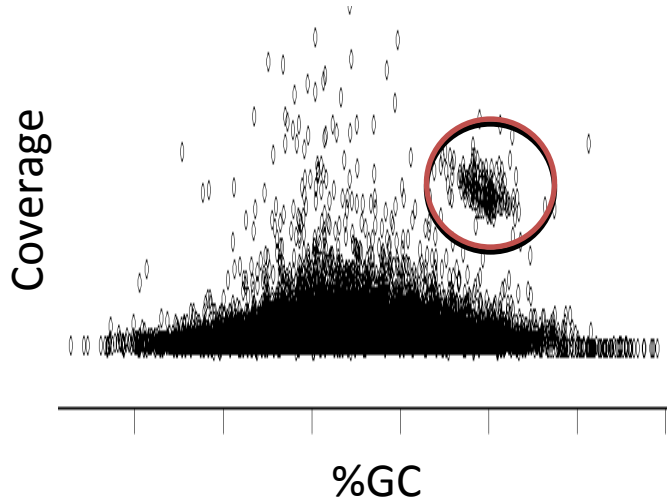
2.5 kbp jumping lib proof of principle: Scaffolding the spruce chloroplast



1 run 454 assembled with Newbler => 9 chloroplast contigs
Mapping 1% of 1 lane MP data => 1 circular scaffold

Detected 1 translocated inversion compared to *Picea sitchensis*

Mito

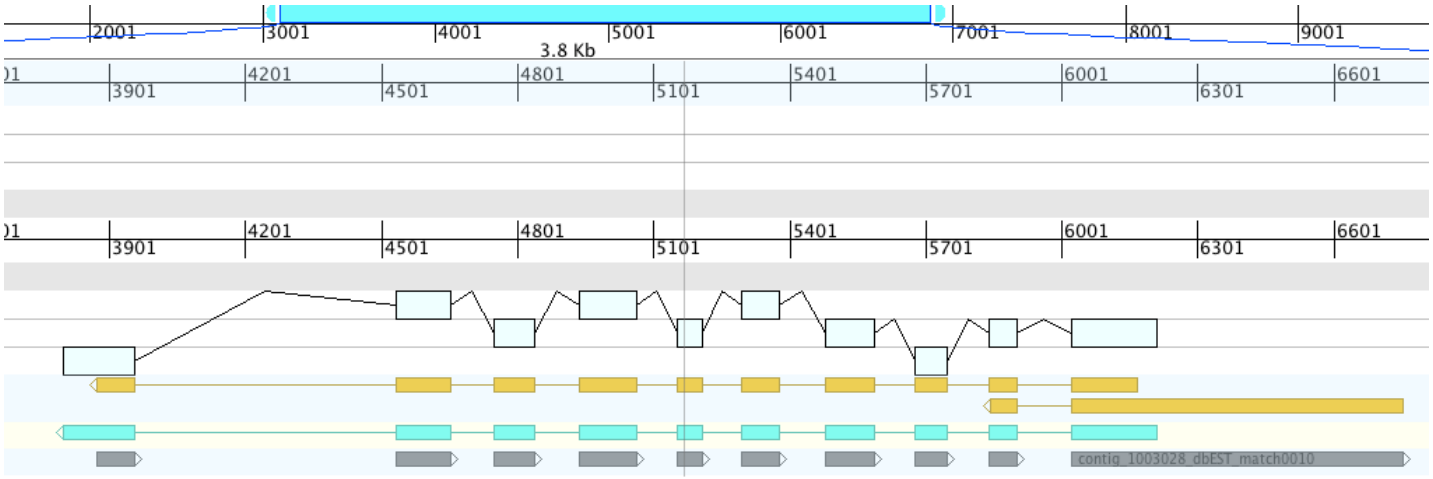


38 of 39 *Cycas* mitochondrial genes found in this set

Potential mito contigs sum up to 5 Mbp (!)
(The *Cycas*mito gene contigs alone sum up 1.4 Mbp)

Much longer contigs than nuclear DNA (less repeats?)
N50 - contigs: 50 kbp
- scaffolds: 289 kbp

Do we have the genes yet?

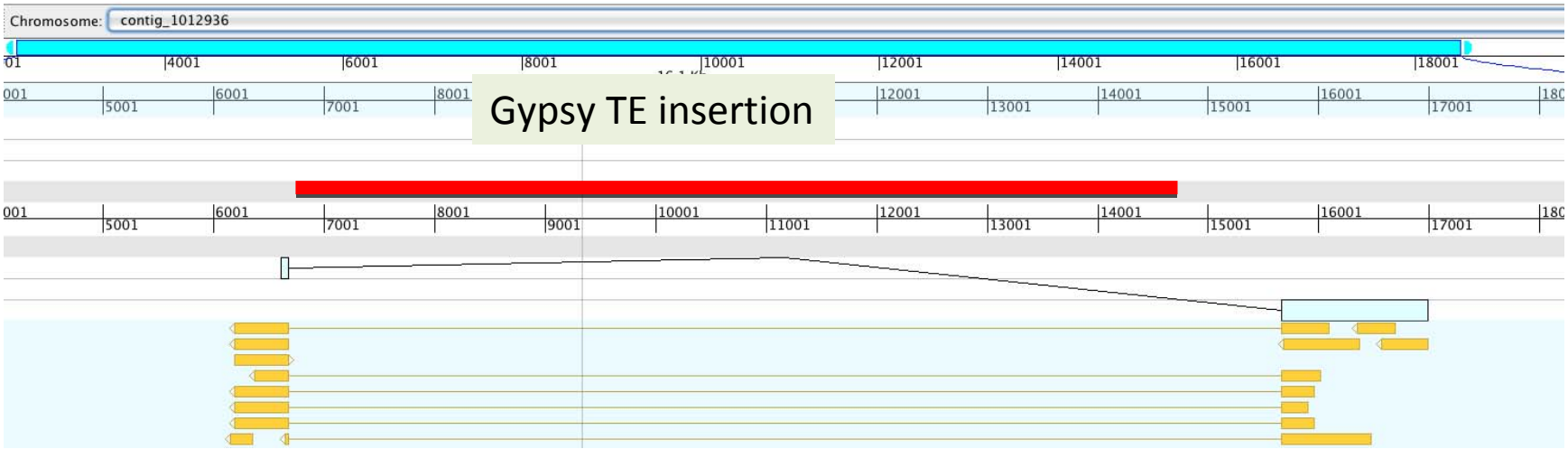


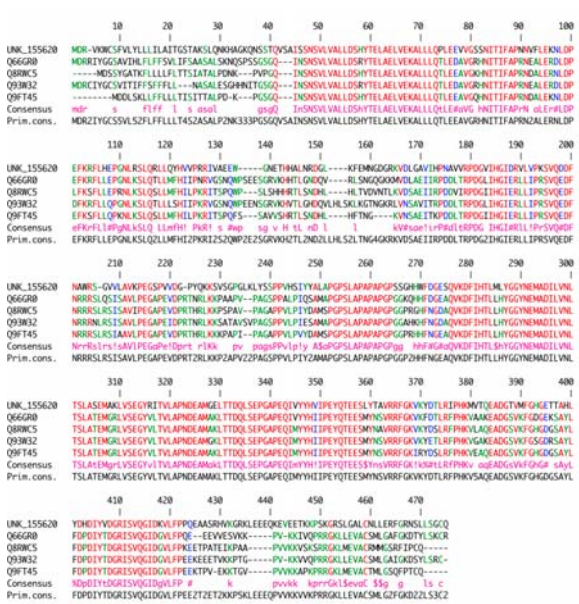
A complete gene structure (2.4 kbp)

Map 27.000 FL-cDNA:s (White spruce) to the WGS

>30% contained in a single contig
(60% well covered but split on multiple contigs)

Long introns



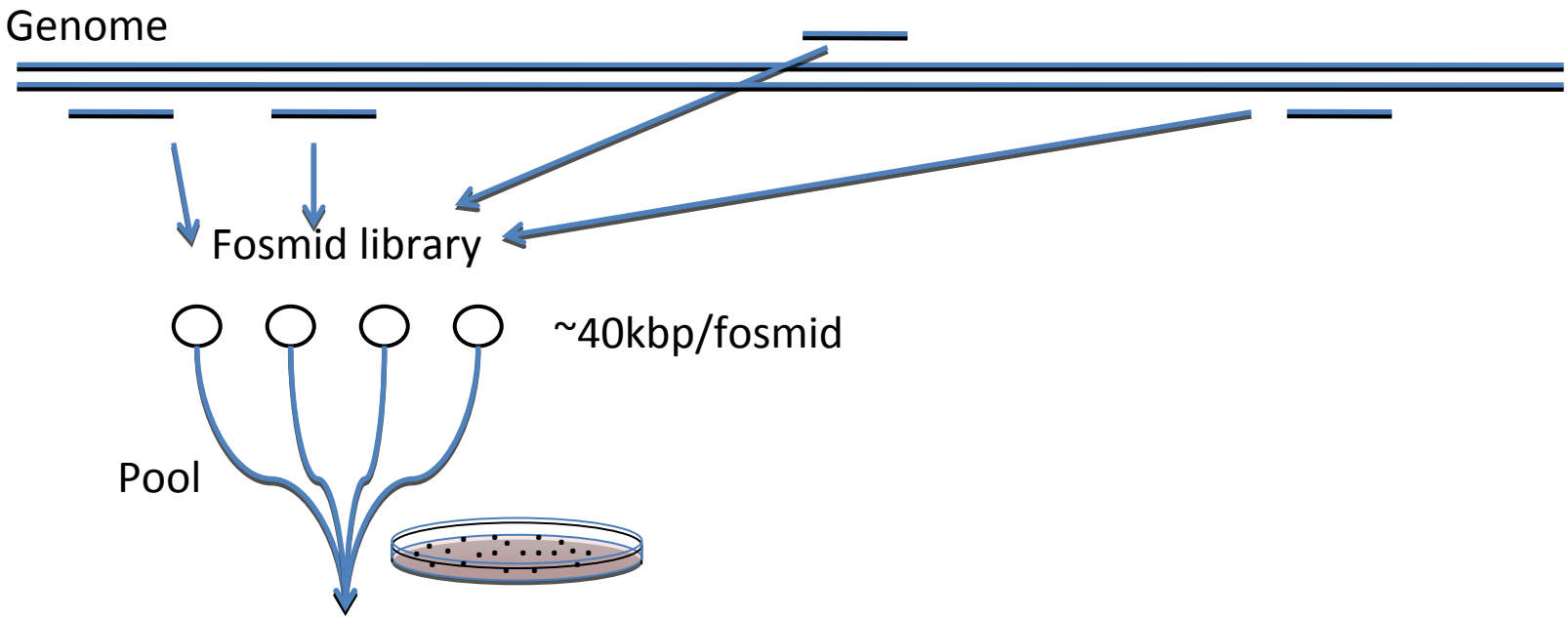


Fasciclin-like arabinogalactan protein (putative cell adhesion)

8 kbp TE intron insertion

Conserved protein sequence

Fosmid pool strategy



~1000 fosmids/pool (~40 Mbp)
8 pools/HiSeq lane => 60X fosmid coverage
500 pools (~8flowcells) => 1X genome coverage

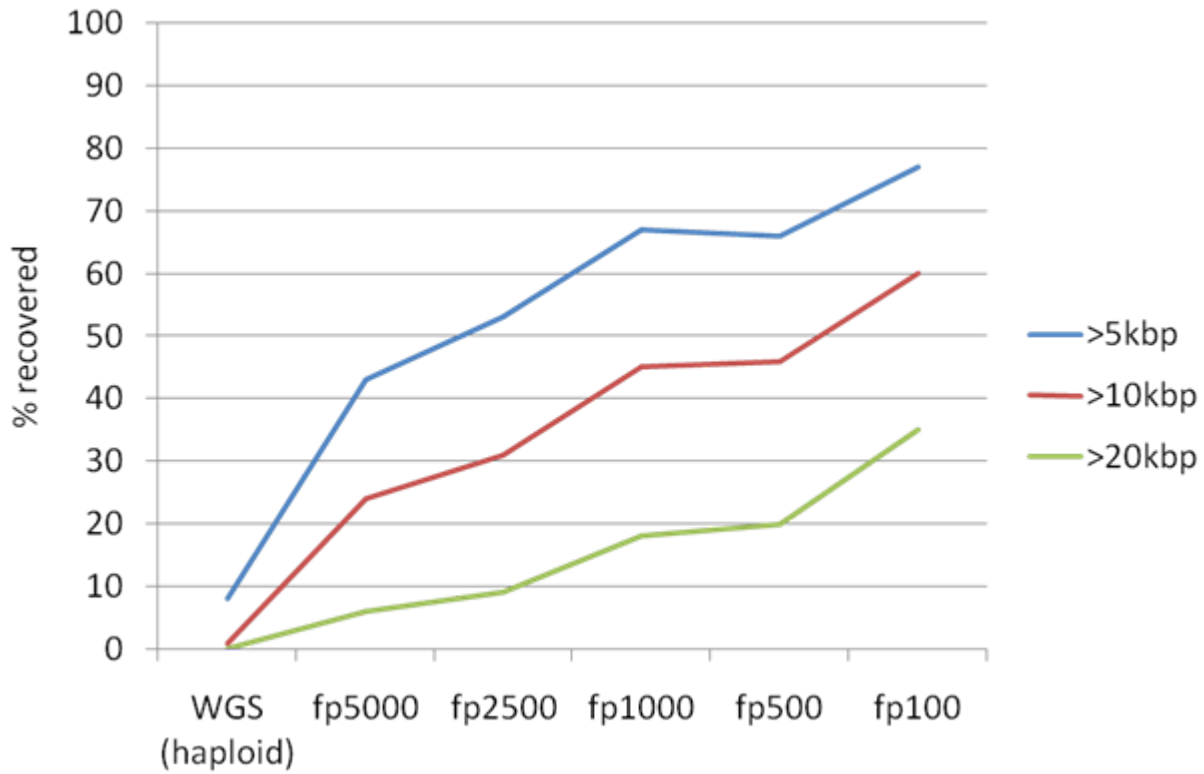
↓

500 assemblies á 40 Mbp

↓

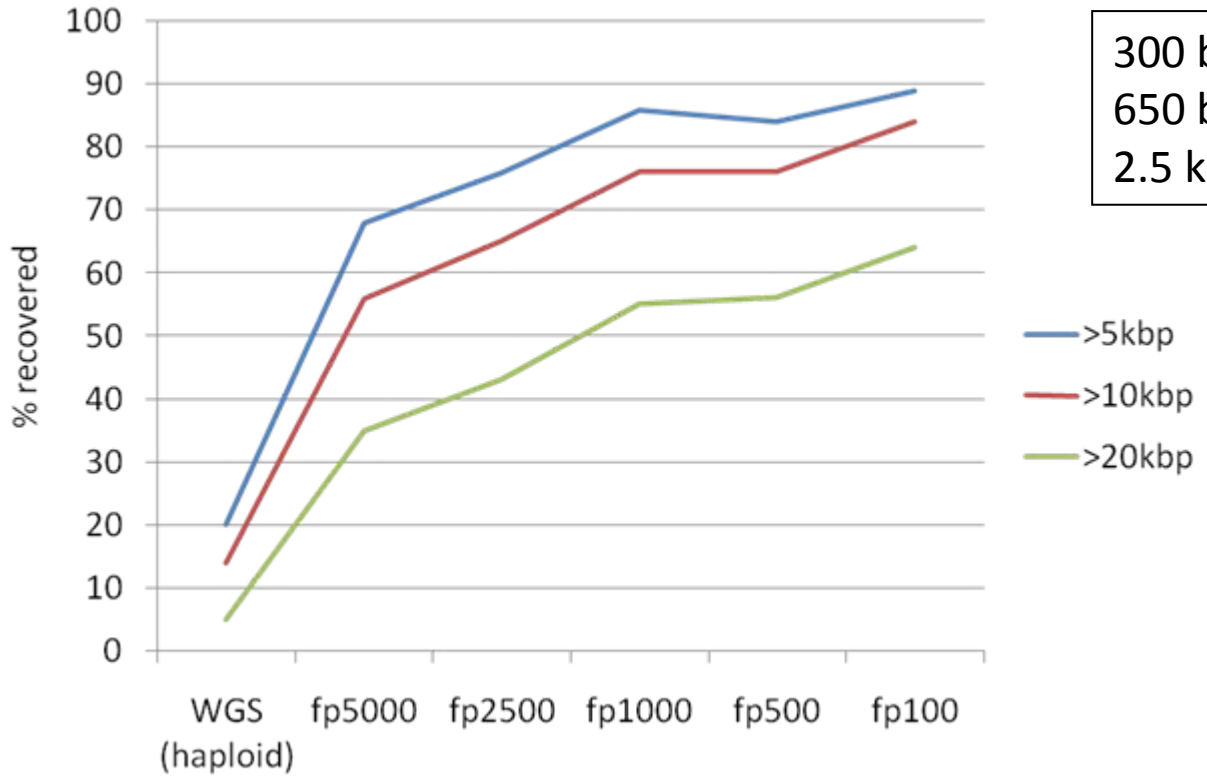
Merge all assemblies (including WGS)

Fosmid pool size



20 kbp contigs:
3 pools of 1000 fosmids (<1% of the genome) enough to beat WGS…

Fosmid pool size (scaffolded)



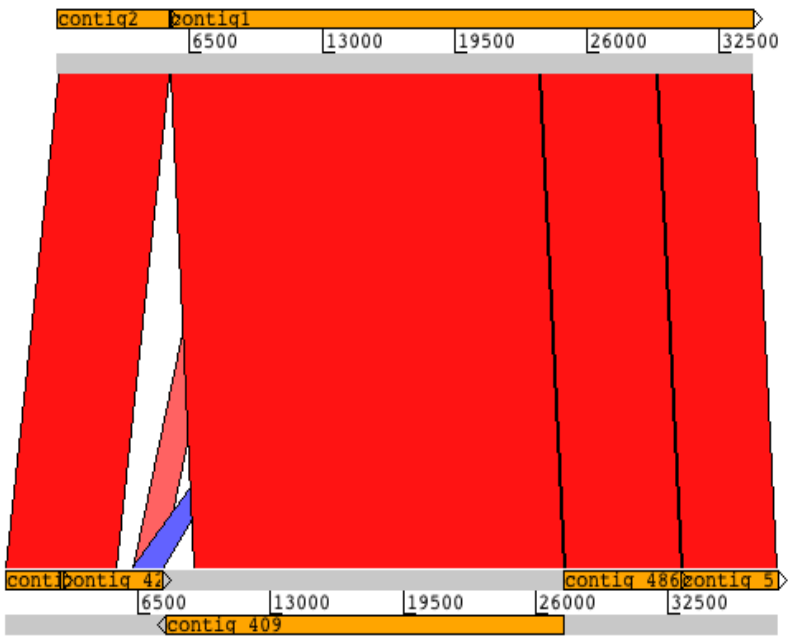
300 bp lib (pool)
650 bp lib (WGS)
2.5 kbp lib (WGS)

Assembly/scaffolding validation

SciLifeLab

Validation of pool assemblies by individual fosmid assemblies

Pick individual fosmid from known pool → PacBio 454 (Illumina) → Assembly (Finishing) → Benchmark pool



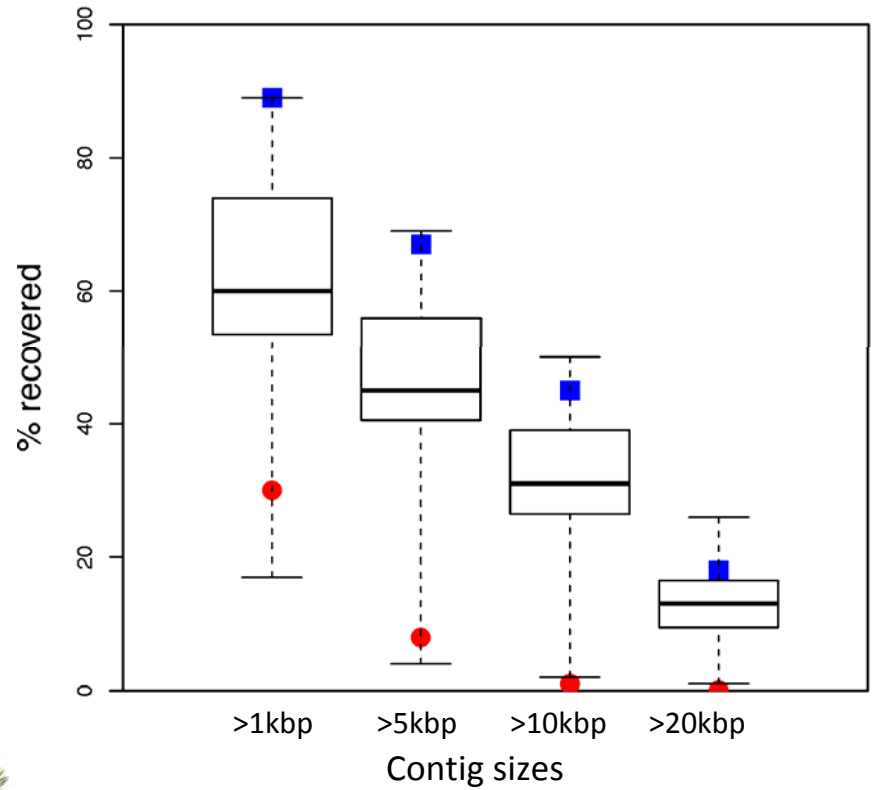
A single fosmid from the fp100 pool
(PacBio: 2 contigs, 34kbp)

The corresponding scaffold from the same pool
(CLCbio + BESST: 5 contigs in 1 scaffold, 38kbp)

Fosmid pools: Scaling up

SciLifeLab

Fosmid pools	**Status**
Pool size trials	5 pools done
First 500 pools (1X)	300 libs done (20% analysed)
Additional 1500 pools (3X)	Production in progress



Production mode pools not as great as the trials, but still much better than the WGS

Counts: 700 -1300 fosmids per pool
Coverage: ~30X (aimed for 75X)

Trial pool fp1000
WGS

What's next?

- Complete first 500 fosmid pools (Feb 2012)
- More paired data (10 kbp, fosmid ends)
- Assembly merging
- More individual fosmids for benchmarking
- Assembly error detection methods
  (paired reads, independent datasets)

Tuesday, Lucigen workshop
14.10 Björn Nystedt

“Fosmid pool sequencing of the 20 Gbp genome of Norway spruce *(Picea abies)*”

Scaffolding of WGS

WGS (haploid)	**Status**
PE (150bp, 300bp, 650bp)	(20X)

WGS (diploid)	**Status**
454 (SE)	1.5X
PE (150bp, 300bp, 650bp)	55X
MP (2.5 kbp)	40X span
MP (10kbp)	Trials in progress
Fosmid ends	Trials in progress

Assembly stats, 15X haploid scaffolded

- 31 % (30%) in contigs>1 kbp
- 20 % (8%) in contigs>5 kbp
- 14 % (1%) in contigs > 10 kbp