

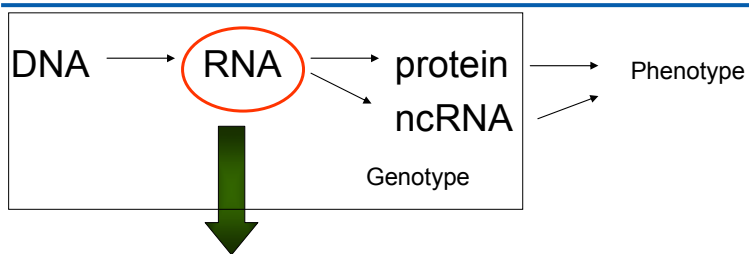


# Transcriptome Analysis of Non-Model Organisms using Short-Read Sequences

Lesley Collins, Patrick Biggs  
Claudia Voelckel, Simon Joly

Allan Wilson Centre,  
Massey University  
New Zealand

## Transcriptome Analysis

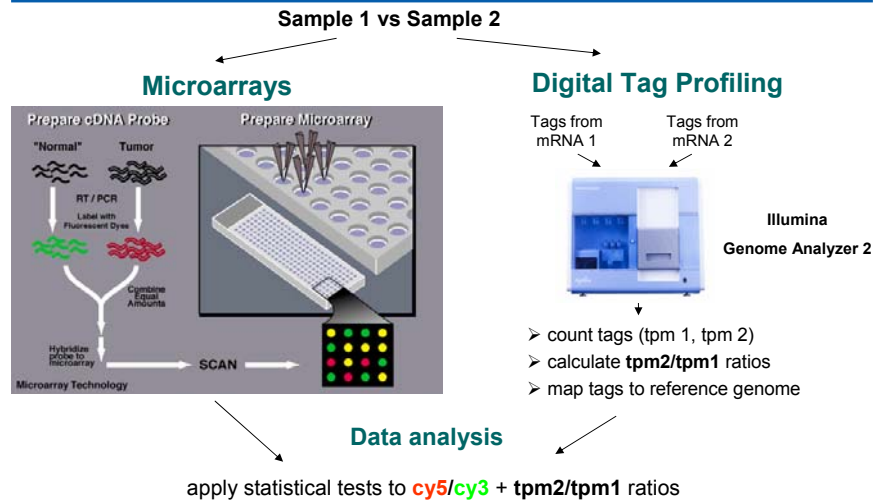


Which genes are expressed in certain situations?

Which genes are expressed differently in different species?

ESTs (Expressed Sequence Tags) are collected from RNA samples reflecting which genes are being expressed.

## Expression studies – Of chips and tags



1 December 2008

GIW 2008 Collins and Biggs

## Benefits and constraints of...

### Microarray studies



- microarray platform needed
- queries limited to genes present on the array
- multi-species comparisons tricky

- 100 ug per sample
- low intensity data unreliable
- intense data pre-processing

- reasonable costs

&

### Tag profiling studies



- open to any organism (but tag annotation depends on available genome information)
- any transcript detectable (restriction site)
- multi-species comparisons straightforward

- 1ug RNA per sample

- single copy transcripts detectable (3 tpm)
- little data pre-processing

- still expensive: cost likely to decrease when multiplexing biological replicates

1 December 2008

GIW 2008 Collins and Biggs

## The *Pachycladon* Transcriptome Project



### Alpine Cress (*Pachycladon*, Brassicaceae)



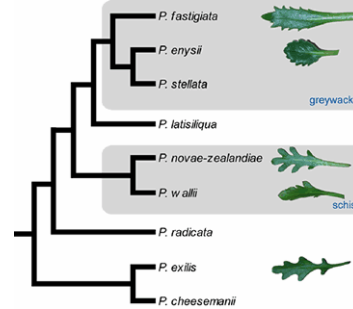
*P. enysii*

2492 m  
obligate greywacke



*P. fastigiata*

1485 m  
can grow on  
schist/greywacke



What are the processes of ecological divergence?

Has there been an adaptive radiation to different geological substrates?

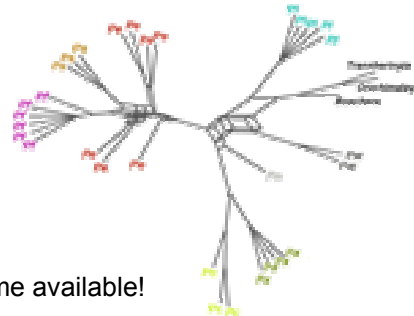
1 December 2008

GIW 2008 Collins and Biggs

## But *Pachycladon* is a non-model plant...

There are 8 South Island species of recent origin (<1 mya):

- P. fastigiata*
- P. enysii*
- P. stellata*
- P. novae-zealandiae*
- P. wallii*
- P. cheesemanii*
- P. exilis*
- P. latisiliqua*



- There is no *Pachycladon* genome available!
- Can we look at the *Pachycladon* transcriptome without having to sequence the entire genome?


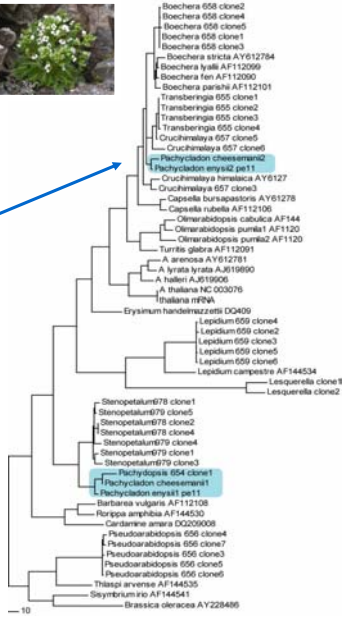
1 December 2008

GIW 2008 Collins and Biggs

## Chalcone Synthase (coding regions only)

# Pachycladon is polyploid!

- Member of *Brassicaceae* family
- Allopolyploid from very diverged *Brassicaceae* parents
- 2 copies for most (all?) genes
  - One from Himalayas
  - Other from Australia
- Rough divergence time estimates from *Arabidopsis*
  - 2 – 10 m year (S. Joly)

**Simon Joly, submitted**

1 December 2008

GIW 2008 Collins and Biggs

## Sequencing of 3 *Pachycladon* species

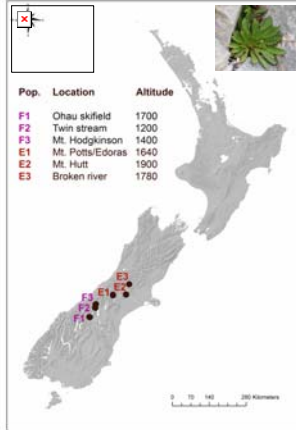
Platform	Template	Sequence length	<i>P. enysii</i>	<i>P. fastigiata</i>	<i>P. cheesemanii</i>
Solexa	cDNA	36 bp	X		X
454	cDNA	40-320 bp		X	
Solexa	mRNA tags	18 bp	X	X	

1. Construct EST libraries with different sequencing depth using the same platform
2. Compare platforms for EST library construction
3. Establish EST databases/Reference transcriptomes
4. Use EST libraries for development of molecular markers
5. Digital Gene Expression pilot study to be compared with earlier microarray results

1 December 2008

GIW 2008 Collins and Biggs

## *P. enysii* and *P. fastigiata* – Sampling



Claudia Voelckel and Peter Heenan

1 December 2008

GIW 2008 Collins and Biggs

## *Some people had much more fun!*

Pete Lockhart collecting samples from Molesworth station!!



1 December 2008

GIW 2008 Collins and Biggs

## *Pachycladon* transcriptome analysis

- Collect reads and assemble into contigs → Pachy spp. ESTs
  - 454 and Solexa
- Compare contigs/reads against Arabidopsis
  - Which genes are easily identified?
- Map tags against Arabidopsis
  - Preliminary expression analysis
- Annotate contigs into 'potential ESTs' and map tags against these *Pachycladon* 'ESTs'
  - *Pachycladon* Expression Analysis !!!!

1 December 2008

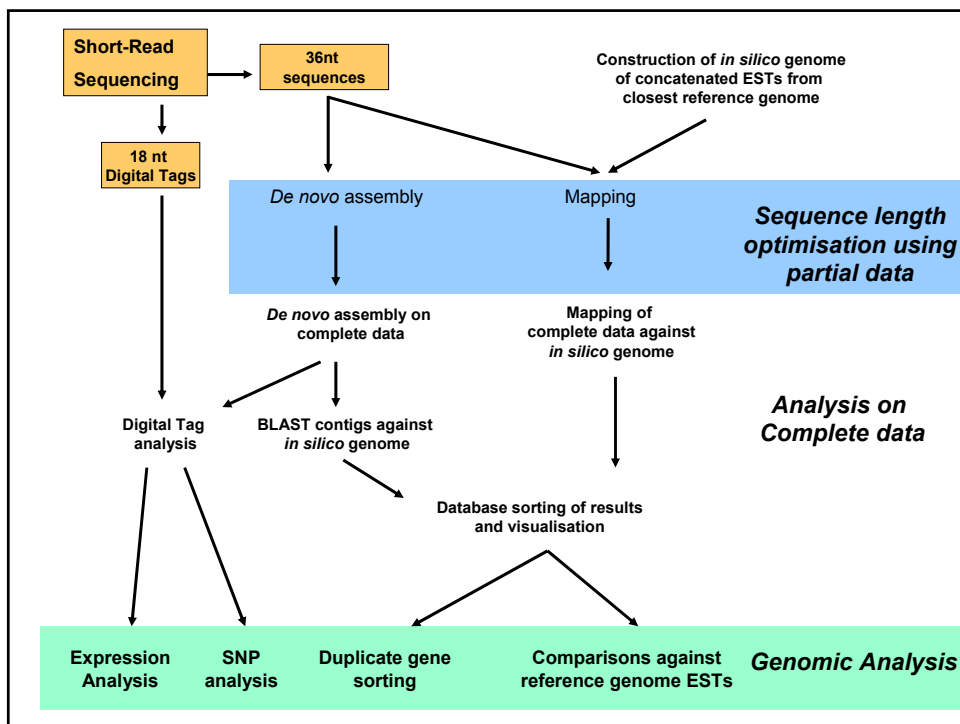
GIW 2008 Collins and Biggs

## *Pachycladon* data collected

- 63,870 FLX-454 reads for *P. fastigiata*
    - 78 contigs (length > 500 nt)
    - 3825 contigs (length > 100 nt)
  - 40 million Solexa 36nt reads for *P. enysii*
  - 10 million Solexa 36nt reads for *P. cheesemanii*
  - 14.9 million 18 nt Digital tags for *P. enysii*
  - 15.8 million 18 nt Digital tags for *P. fastigiata*
- Otago Genomics Sequencer, Otago University, New Zealand
- ALLAN WILSON CENTRE GENOME SERVICE
- AWC Genome Sequencing Service, Massey University, New Zealand

1 December 2008

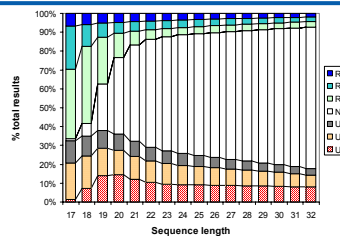
GIW 2008 Collins and Biggs



## Optimised mapping against *Arabidopsis* TAIR7 ESTs

Data was mapped with Eland (Illumina)

- Allows only 2 mismatches per sequence
- Not all data will map
  - Evolutionary distance between *Pachycladon* and *Arabidopsis*



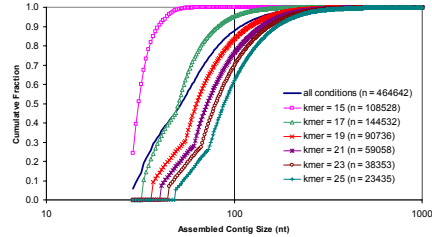
ELAND result	ELAND Result type	Length = 19 (1 lane)	% Total (1 lane)	Number (7 lanes)	Av % Total (7 lanes)	Std Dev (7 lanes)
U0	Unique 0 mismatch	341,804	8.99	3,577,678	8.86	1.2
U1	Unique 1 mismatch	543,248	14.29	5,697,913	14.23	0.34
U2	Unique 2 mismatch	520,442	13.69	5,547,612	13.85	0.29
R0	Repeat 0 mismatch	960,119	25.26	1,943,961	24.84	0.93
R1	Repeat 0 mismatch	191,895	5.05	2,914,570	4.88	0.22
R2	Repeat 0 mismatch	275,036	7.24	10,455,270	7.27	0.16
QC	Quality filtered	14	0.00	121	0.00	0.00
NM	No match	968,842	25.49	9916563	26.07	0.79
<b>Total</b>	-	<b>3,801,400</b>	<b>100</b>	<b>40,053,567</b>	<b>100</b>	-

1 December 2008

GIW 2008 Collins and Biggs

## Optimised Contig assembly parameters

- Edena
  - Hernandez et al. 2008 Genome Res 18:802
- Velvet
  - Zerbino and Birney 2008 Genome Res 18:821

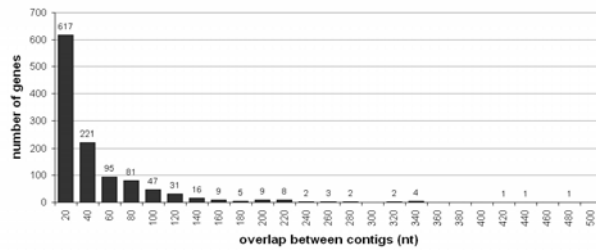


K-mer Size	Number of Contigs	BLAST hits	Number of Contigs with BLAST hits	% Contigs with BLAST hits	Number of genes hit		
					All aligns	Aligns >40 nt	Aligns >85 nt
15	2	3	1	50.00	3	2	0
17	12,638	18,844	8,780	69.50	10,579	9,312	6,786
19	22,631	38,535	16,413	72.50	14,906	13,105	10,304
21	20,531	39,554	15,227	74.20	14,594	12,267	9,504
23	16,873	34,486	12,739	75.50	13,209	10,732	8,222
25	12,750	27,360	9751	76.50	10,871	8,852	6,595
<b>Total</b>	<b>85,425</b>		<b>62,911</b>	<b>73.65</b>			

1 December 2008

GIW 2008 Collins and Biggs

## Optimised contig mapping parameters



- Number of *Arabidopsis* ESTs to which *Pachycladon* contigs mapped with any overlap (bin size = 20 nt)
- E.g. These contigs mapped to 617 *Arabidopsis* ESTs with an overlap of 1-20 nt

1 December 2008

GIW 2008 Collins and Biggs



## Arabidopsis TAIR7 vrs TAIR8



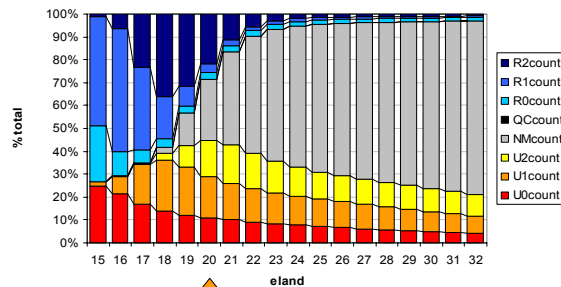
- This project began using TAIR7...  
...then TAIR8 was released
- TAIR7
  - 27,029 protein coding genes (3866 have splice variants - 12%)
- TAIR8
  - 27,235 protein coding genes (4330 have splice variants - 13%)
  - 1291 new genes
  - 2009 new gene models
  - 23% of existing TAIR7 genes modified

1 December 2008

GIW 2008 Collins and Biggs

## Mapping against the *Arabidopsis* TAIR8 Genome

- **Arabidopsis transcriptome:** 38,963 ESTs (27,235 protein coding genes)
- With Eland reads map uniquely (U), repeatedly (R) or not at all (NM & QC)



Mapping results for  
1 lane (3.8 million reads)  
with 20mers:

NM	27%
R	29%
U	45%

↑ Optimal k-mer: 20

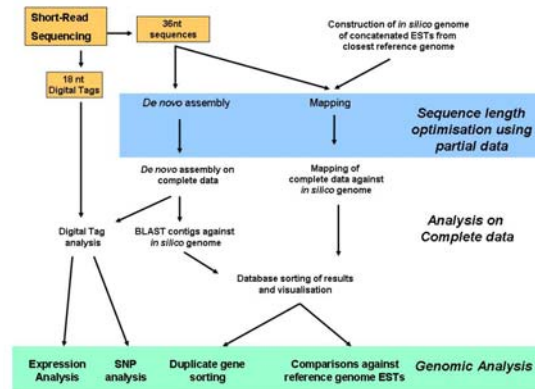
Conclusion: TAIR8 is slightly better but TAIR7 data still very useful

1 December 2008

GIW 2008 Collins and Biggs

## Now onto the full analysis

- Mapping of complete data to TAIR8
- De novo assembly of Pachycladon ESTs
- Analysis of Digital Tags



1 December 2008

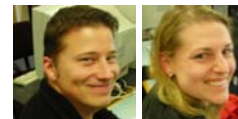
GIW 2008 Collins and Biggs

## Contigs assembled using Velvet and Edena

- Criteria
  - BLASTN HSP alignment must cover at least 90% of the contig length
  - Alignment must contain at least 90% identical nucleotides

	<i>P. ensyia</i> Edena	<i>P. ensyia</i> Velvet	<i>P. cheesemanii</i> Edena	<i>P. cheesemanii</i> Velvet
Contigs >500nt	542	1042	365	573
Hits in Arabidopsis	97	125	68	72

- Contigs from 454 for *P. fastigiata* reanalysed
- EST database Version 0.1 under construction
- Many thanks to Oliver Deusch and Nicole Grünheit



1 December 2008

GIW 2008 Collins and Biggs

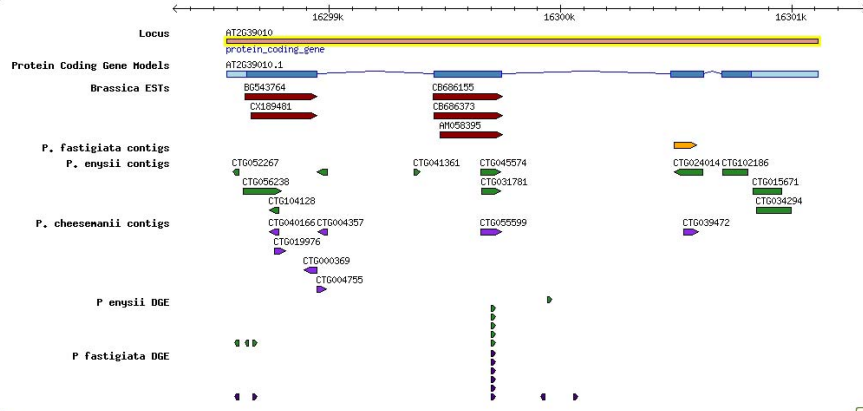


# Arabidopsis thaliana Genome with Pachycladon spp. data

Running on K9 at AWC  
Modified from this [TAIR8 Database release](#)

Showing 3 kbp from Chr2, positions 16,298,324 to 16,301,323

- [Instructions](#)
- [\[Show banner\]](#) [\[Bookmark this\]](#) [\[Link to Image\]](#) [\[High-res Image\]](#) [\[Help\]](#) [\[Reset\]](#)
- [Search](#)
- [Overview](#)
- [Details](#)



1 December 2008

GIW 2008 Collins and Biggs

## Gbrowse against Arab Tair8

Home Help Contact About Us Login/Register

Gene

---

Search
Browse
Tools
Stocks
Portals
Download
Submit
News

**Locus: AT2G39010**

Date last modified: 2003-05-02

TAIR Accession: Locus:2064885

Representative Gene Model: [AT2G39010.1](#)

Gene Model: [protein\\_coding](#)

Type: protein\_coding

Other names: PIP2.6, PIP2E, PLASMA MEMBRANE INTRINSIC PROTEIN 2.6, PLASMA MEMBRANE INTRINSIC PROTEIN 2E, T7F6.18, T7F6\_18

Description: PIP2.6/PIP2E (plasma membrane intrinsic protein 2.6), water channel. Identical to Probable aquaporin PIP2-6 (PIP2-6) [Arabidopsis thaliana] (GB:Q9ZV07); similar to PIP2.5/PIP2D (plasma membrane intrinsic protein 2.5), water channel [Arabidopsis thaliana] (TAIR:AT3G54820.1); similar to water channel protein [Brassica rapa] (GB:ABL97985.1); contains InterPro domain Aquaporin, (InterPro:IPR012269); contains InterPro domain Major intrinsic protein, (InterPro:IPR000425)

Map Detail Image

Annotations	Category	Relationship Type	Keyword
	GO Biological Process	involved in	response to nematode, transport
	GO Cellular Component	located in	membrane
	GO Cellular Component	located in	membrane, plasma membrane
	GO Molecular Function	has	water channel activity
	Plant structure	expressed in	root
	Plant structure	expressed in	leaf, cultured cell

1 December 2008

GIW 2008 Collins and Biggs

## Finding Pachycladon gene pairs

Most genes will have two copies when compared to Arabidopsis

```

AT5G64740. : AATGGGAAAGGAGATTGGGATCTATGGTTCTCTTACCSAAGATATTCTACGGGTTTAAATGCAT : 72
19_289547 : -----GAGATATTCTACGGGTTTAAATGCAT : 29
19_289442 : -----GAGATTGGATGGATCTATGGTTCTCTGACASAGATATTCTACGGGTTTAAATGCATG : 61
                gagattgg tggatctatggttctgt ac gaAGATATTCTtACgGGtTTcAAgATGCATt

AT5G64740. : CTCATGGTTGGAGATCTGTTTATGTACACCAGGTTGCGGCITTCAAAGGATCGCTCCATCAATCTTT : 144
19_289547 : CTCATGGTTGGAGATCTGTTTATGTACACCAGGTTGCGGCITTCAAAGGATCGCTCCATCAATCTTT : 101
19_289442 : CTCATGGTTGGAGATCTGTTTATGTACACCAGGTTGCGGCITTCAAAGGATCGCTCCATCAATCTTT : 133
                CtCaTGGtTGGAGaTctgTtTAtTgTAcacCaAAGtTaccGGcITTCAAAGGaTcAgCtCCaATCAATCTTT

AT5G64740. : CCGATCGTCTCCATCAAGTTCTTCGATGGCGCTTGGGTCGGTTGA : 190
19_289547 : CCGATCGTCTCCATCAAGTTCTTCGATGG----- : 130
19_289442 : CCGATCGTCTCAA----- : 146
                CaGAtCGTCTccAtcaagtTcttcgatgg
    
```

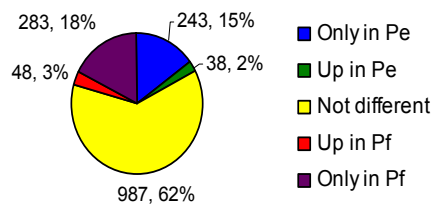
E.g. Two overlapping *Pachycladon* contigs (114bp) that map to At5g64740.1, cellulose synthase 6

1 December 2008

GIW 2008 Collins and Biggs

## Digital Tag Analysis: *P. enysii* and *P. fastigiata* against Arabidopsis

- Pe - 5,591,951 sequences (37.64%) mapped uniquely
- Pf - 5,767,796 sequences (36.48%) mapped uniquely
- 1/3 of all sequences mapped repeatedly
- 1/3 of all sequences did not map



### *P. enysii*

Genes	U0	U012
Mapped	15296	31375
tpm > 3	1311	19451
Occur in both species	1068	18263
Occur only in <i>P. enysii</i>	243	1188

### *P. fastigiata*

Genes	U0	U012
Mapped	16204	31511
tpm > 3	1351	20753
Occur in both species	1068	18263
Occur only in <i>P. fastigiata</i>	283	2490

Lydia Hopp and Janina Mothes

1 December 2008

GIW 2008 Collins and Biggs

## *P. enysii* vs. *P. fastigiata* – Previous Chip study



*P. fastigiata* (1485 m, glabrous)

310 genes ↑



*P. enysii* (1885 m, hairy)

324 genes ↑



### In press: *Molecular Ecology*!

Transcriptional and biochemical signatures of divergence in natural populations of two species of New Zealand alpine *Pachycladon*

C. Voelckel<sup>1\*</sup>, P. B. Heenan<sup>2</sup>, B. Janssen<sup>3</sup>, M. Reichelt<sup>4</sup>, K. Ford<sup>2</sup>, R. Hofmann<sup>5</sup>, P. J. Lockhart<sup>1</sup>

- Gene expression differences predict differences in secondary metabolites (e.g. type of mustard oils produced)
- Follow up studies confirm: *P. enysii* leaves less toxic
- herbivory important for species differentiation !

We can now compare our Digital Tag data with this chip study!

1 December 2008

GIW 2008 Collins and Biggs

## Summary

- *Pachycladon* provides many challenges for genomics:
  - Non-model plant
    - Using *Arabidopsis* as a reference genome worked to a large extent
    - But we need to construct a *Pachycladon* EST database for more meaningful evolutionary studies
  - Polyploid
    - Multiple copies of the same gene
    - Causes problems with contig assembly
  - Alternative splicing
    - Multiple transcripts of each gene copy
    - Can cause problems with mapping



1 December 2008

GIW 2008 Collins and Biggs

## And still more to come...

- Preliminary *Pachycladon* EST libraries in assembly
  - *Pachycladon enysii*
  - *Pachycladon fastigiata*
  - *Pachycladon cheesemanii*
- Comparison of Digital Tag data to *Arabidopsis* Chip study
- Analysis of Digital Tag data with *Pachycladon* ESTs
- SNP analysis of *Pachycladon* data
  - Between *Pachycladon* species
  - Between *Pachycladon* and *Arabidopsis*



1 December 2008

GIW 2008 Collins and Biggs

## Acknowledgements



- Patrick Biggs
- Claudia Voelckel
- Simon Joly
- Pete Lockhart
- Oliver Deusch and Nicole Grünheit (Bill Martin's lab, Duesseldorf)
- Lydia Hopp and Janina Mothes (Greifswald University)
- AWC Genome Sequencing Service
  - Lorraine Berry
  - Maurice Collins
- Allan Wilson Centre (David Penny)
- Institute of Molecular BioSciences
- Marsden Fund
- Royal Society of New Zealand
- Humboldt Foundation



1 December 2008

GIW 2008 Collins and Biggs