Using genomic techniques to investigate interspecific hybridisation and genetic diversity in *Senecio*

Matt Hegarty, Aberystwyth University

NATURAL ENVIRONMENT RESEARCH COUNCIL

Sefydliad y Gwyddorau Biolegol, Amgylcheddol a Gwledig
IBERS ABERYSTWYTH
Institute of Biological, Environmental and Rural Sciences

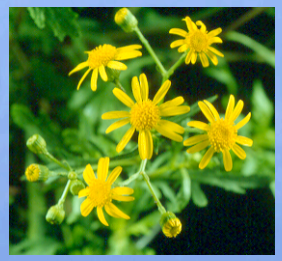
Why are hybridisation and polyploidy important?

- Hybridisation can serve as a source of genetic novelty.
- This can result in adaptive divergence and thus speciation.
- It is predicted that ~70% of higher plants have undergone at least one round of genome duplication over their evolutionary history.
- Approximately 15% of speciation events involve a change in ploidy.
- It is estimated that the majority of polyploid species are the result of interspecific hybridisation.
- Many of the world's most successful crop species are polyploids (bread wheat, coffee, cotton, sugarcane, maize) and often significantly outperform their diploid relatives.

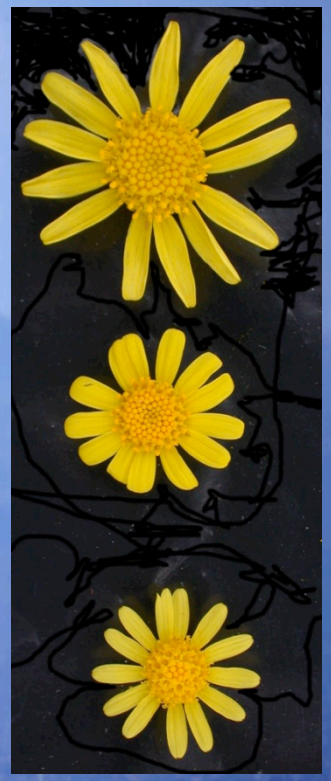
Hybrid speciation in *Senecio*: the homoploid origin of *Senecio squalidus*

Senecio aethnensis
2n = 20

X



Senecio squalidus
2n = 20



Senecio chrysanthemifolius
2n = 20

Abbott *et al.* (2000)
Watsonia 23: 123-138

Hybrid speciation in *Senecio*: the allopolyploid origin of *S. cambrensis*

Senecio squalidus
Diploid, SI

X

Senecio vulgaris
Tetraploid, SC

Senecio x baxteri
Triploid, sterile

Chromosome doubling

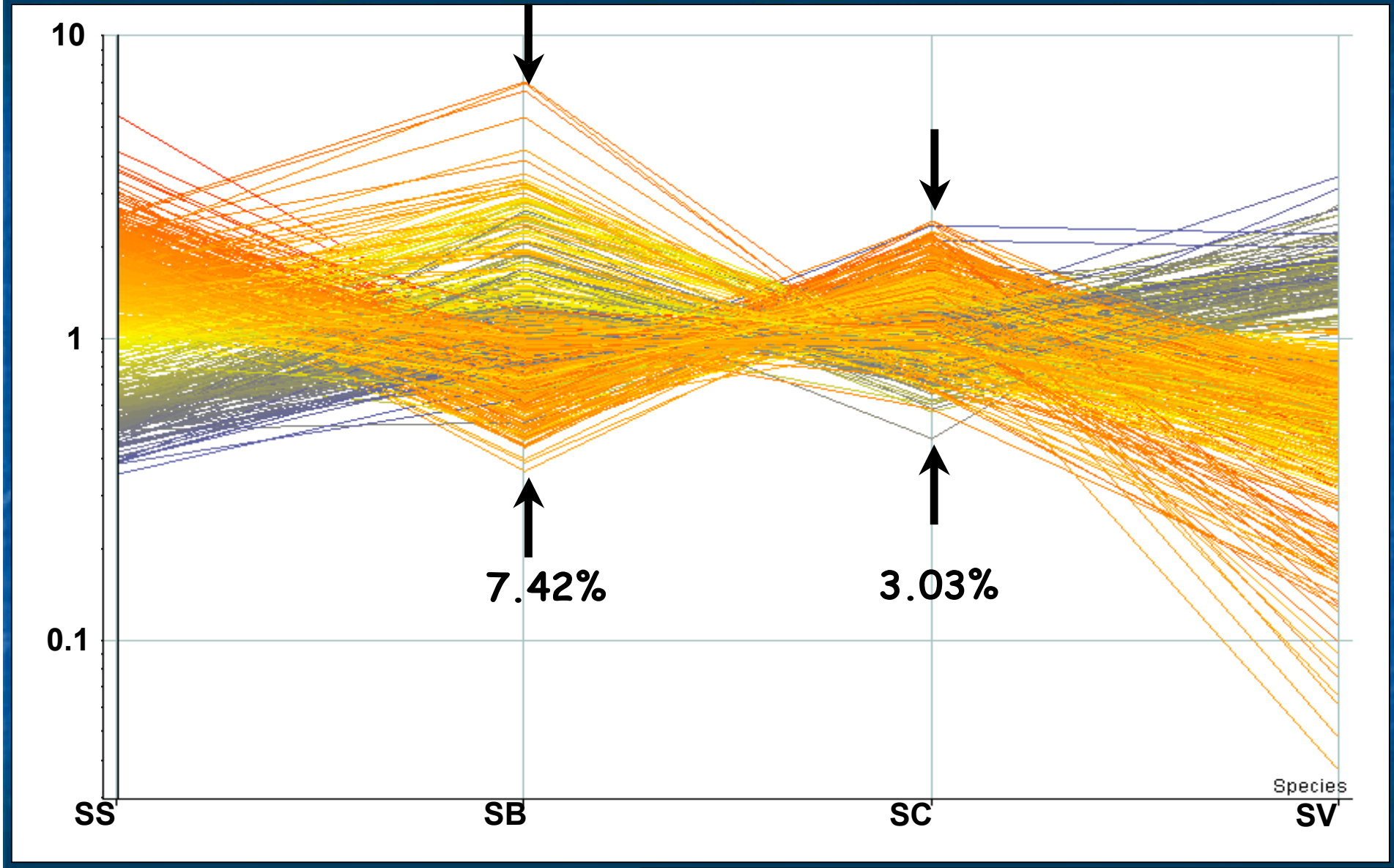
Known to have occurred at least TWICE

Senecio cambrensis
Hexaploid, fertile, SC

Previous work - transcriptomic analysis of interspecific hybrids

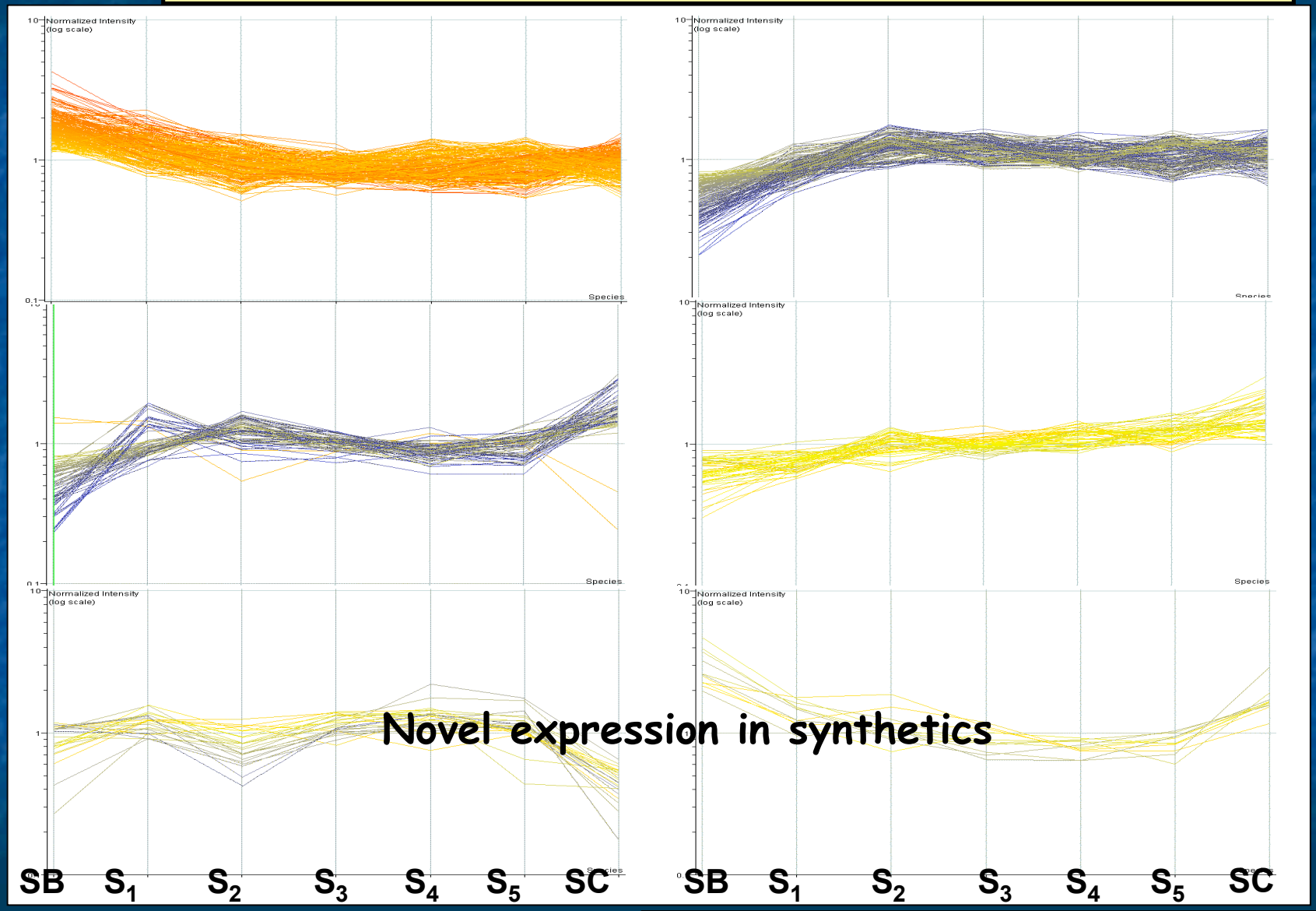
- Constructed cDNA libraries from capitulum and mature flower buds for each of the parental and hybrid species (excl. *S. x baxteri*).
- 1056 clones per library used to create custom cDNA microarrays for gene expression comparisons.
- Libraries sequenced via Sanger sequencing - 11K sequences stored at http://www.seneciodb.org.
- Expression analysis performed to compare parents with both natural and resynthesised hybrids.
- Results show extreme changes to gene expression in a nonadditive manner - that is, hybrids are not merely the average of their parents.

Widespread alterations to gene expression



Hegarty *et al.* (2005) Mol. Ecol. 14: 2493-2510.
Hegarty *et al.* (2008) Phil. Trans. Roy. Soc. 363: 3055-3069.

Genome duplication has a distinct, secondary effect on gene expression in allopolyploids



58%

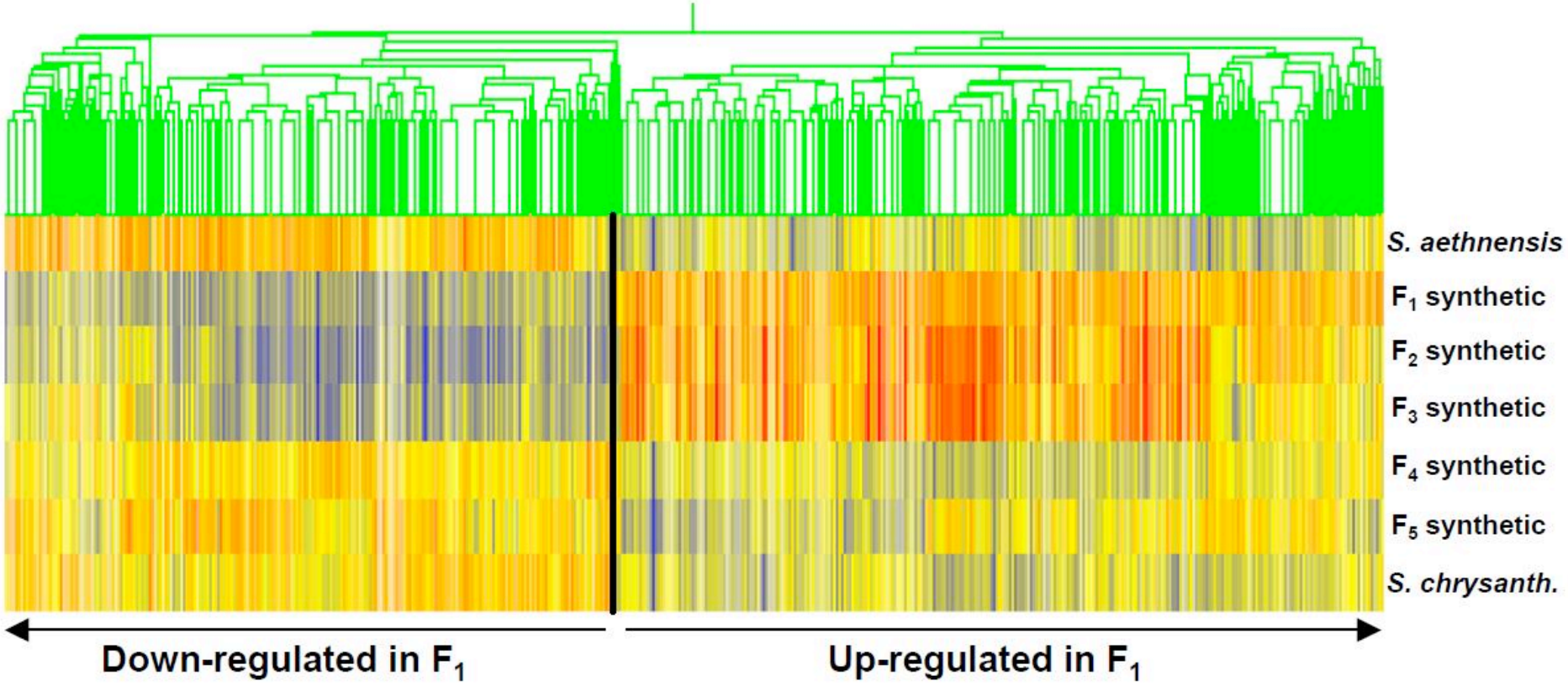
17%

Novel expression in synthetics

Hegarty *et al.* (2006) Current Biology 16: 1652-1659.

Widespread alterations to gene expression

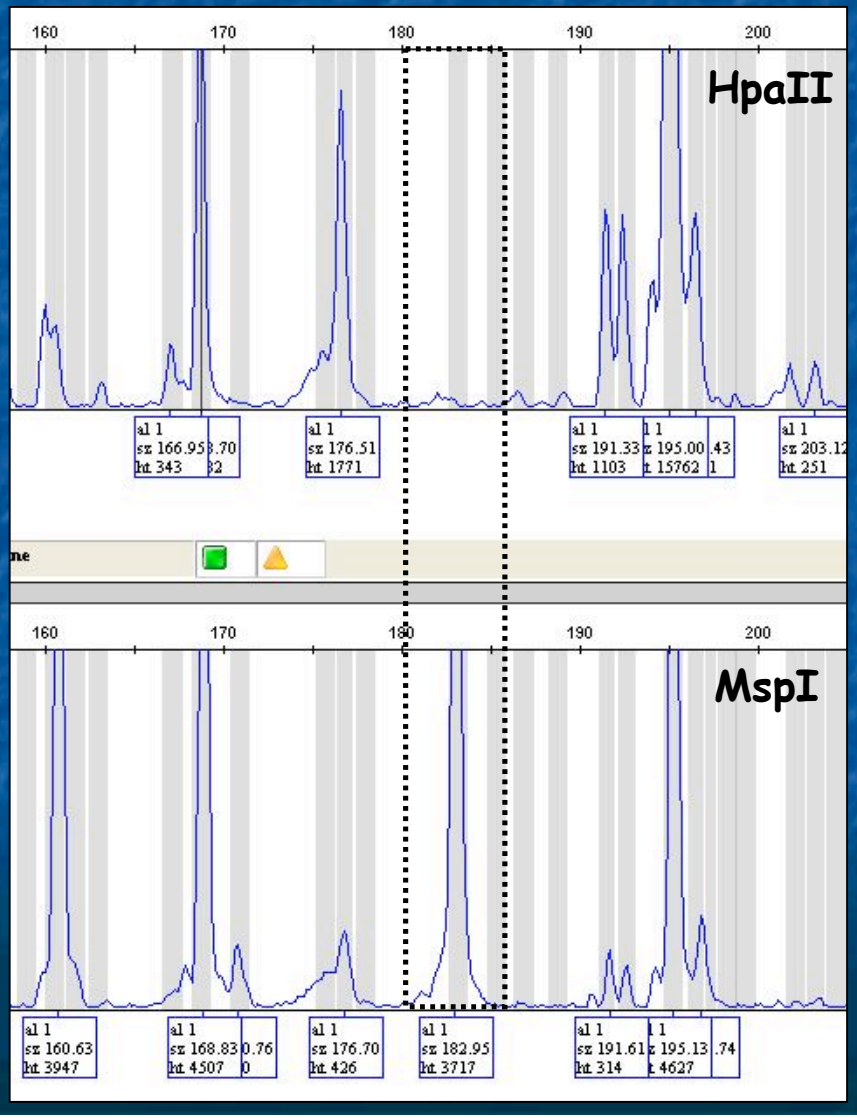
- Same pattern also observed in diploid hybrid *S. squalidus* (both in natural and synthetic hybrids).



5% of array features nonadditive in natural hybrid, up to 10% in first 3 generations of resynthesised hybrids.

Investigating DNA methylation in *Senecio* using MSAP

- MSAP = modified AFLP using EcoRI and either HpaII or MspI.
- If a locus is methylated will see a shift in AFLP profile.



In this example the peak in the MspI sample but not HpaII indicates methylation at the *CCGG* recognition sequence of this locus.

Analysed data from a comparison of three synthetic lines of *S. x baxteri* and two generations of their allohexaploid derivatives plus the parental taxa.

408 loci surveyed, 264 polymorphic between parental taxa.

Methylation changes in *Senecio* polyploids

- Three independent triploid lines of *S. x baxteri* display similar degrees of nonadditive methylation (13.4% on average).

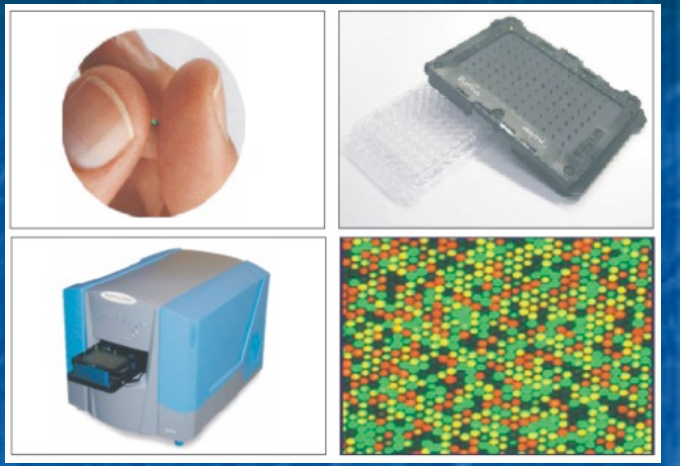
- S_0 allohexaploids derived from triploids are also very similar:

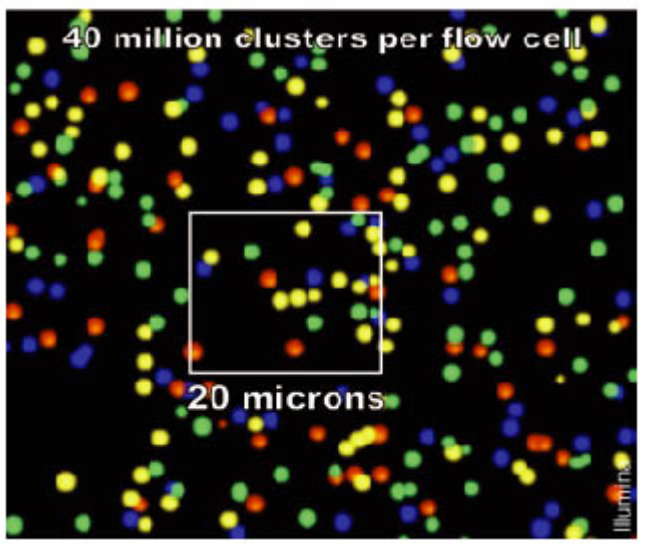
  -73.6% of loci retain nonadditive methylation state seen in triploid.
  -10.1% of loci show change in favoured parental methylation state.
  -2% of loci show novel nonadditive methylation.

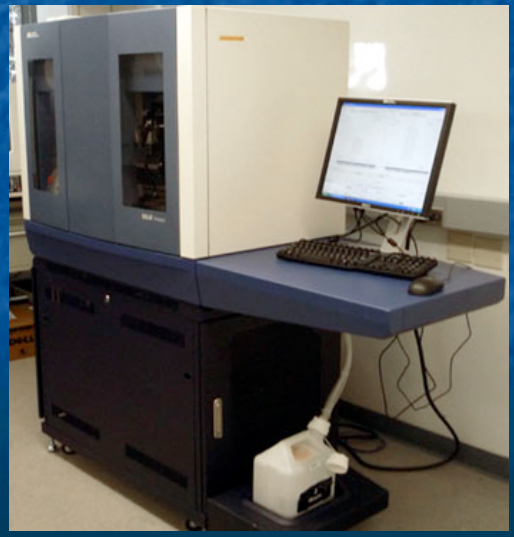
- S_1 allohexaploids also consistent across lines:

  -55.6% of nonadditive methylation seen in triploid is retained.
  -2.5% of loci show novel nonadditive methylation.
  -16.8% of loci differ from S_0 (largely a return to additivity or a change in favoured parental methylation state).

Prospects for using current genomics technology







Prospects for using current genomics technology

- Next-generation sequencing offers a number of techniques which can greatly advance our understanding of genomic effects of hybridisation and polyploidy.
- This technology is also making it simpler to perform assays of genetic diversity on large numbers of markers simultaneously.
- Many of these techniques rely upon availability of at least a partial reference genome, but there are work-arounds.
- We need to think of biological questions that can be addressed using current technology in ways we weren't able to previously.
- All sequence data is valuable! Endpoint assembly of genomes is only possible with availability of community resources.

Studying epigenetics via NGS

- There are a number of epigenetic consequences of interspecific hybridisation and genome duplication that require further study in *Senecio* and in other hybrid systems.

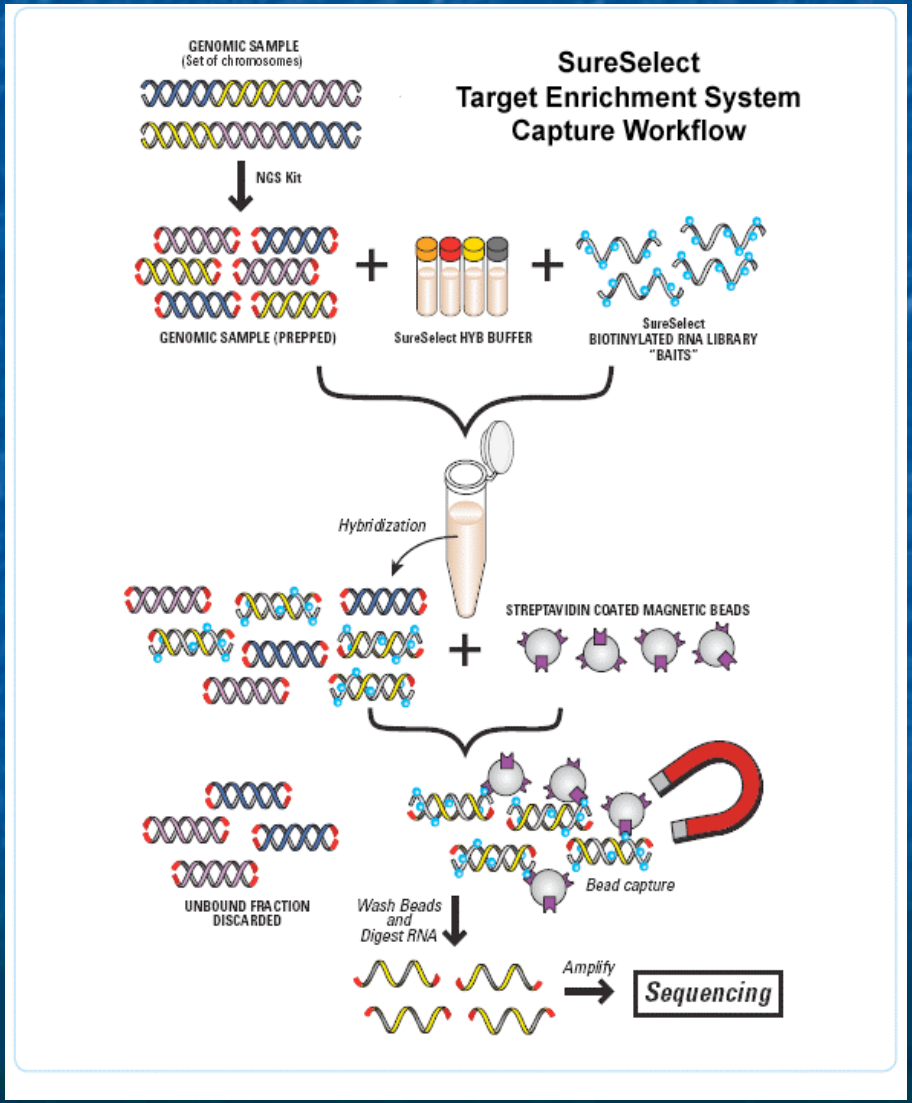
1) Expression bias in polyploids - studies using NGS in cotton have shown that newly formed and natural polyploids display preferential expression of parental gene copies.

2) High-throughput bisulphite sequencing to identify changes in DNA methylation in hybrids/polyploids. Requires a good reference genome.

3) CHiP sequencing of promoter-binding factors to identify alterations in promoter activity in hybrids.

Using sequence capture technology

- Resequencing of known areas of the genome is of use for studying both genetic diversity (SNP/indel detection) and hybridisation:



- For example, can use to detect changes in copy number of target sequences.
- Allows very high coverage of the targets giving confidence in SNP detection and copy # quantitation.
- Enriches for target sequences by 300-7400-fold.
- Since now sequencing a much reduced part of the genome, get sufficient coverage to allow sample multiplexing.

Reduced complexity sequencing

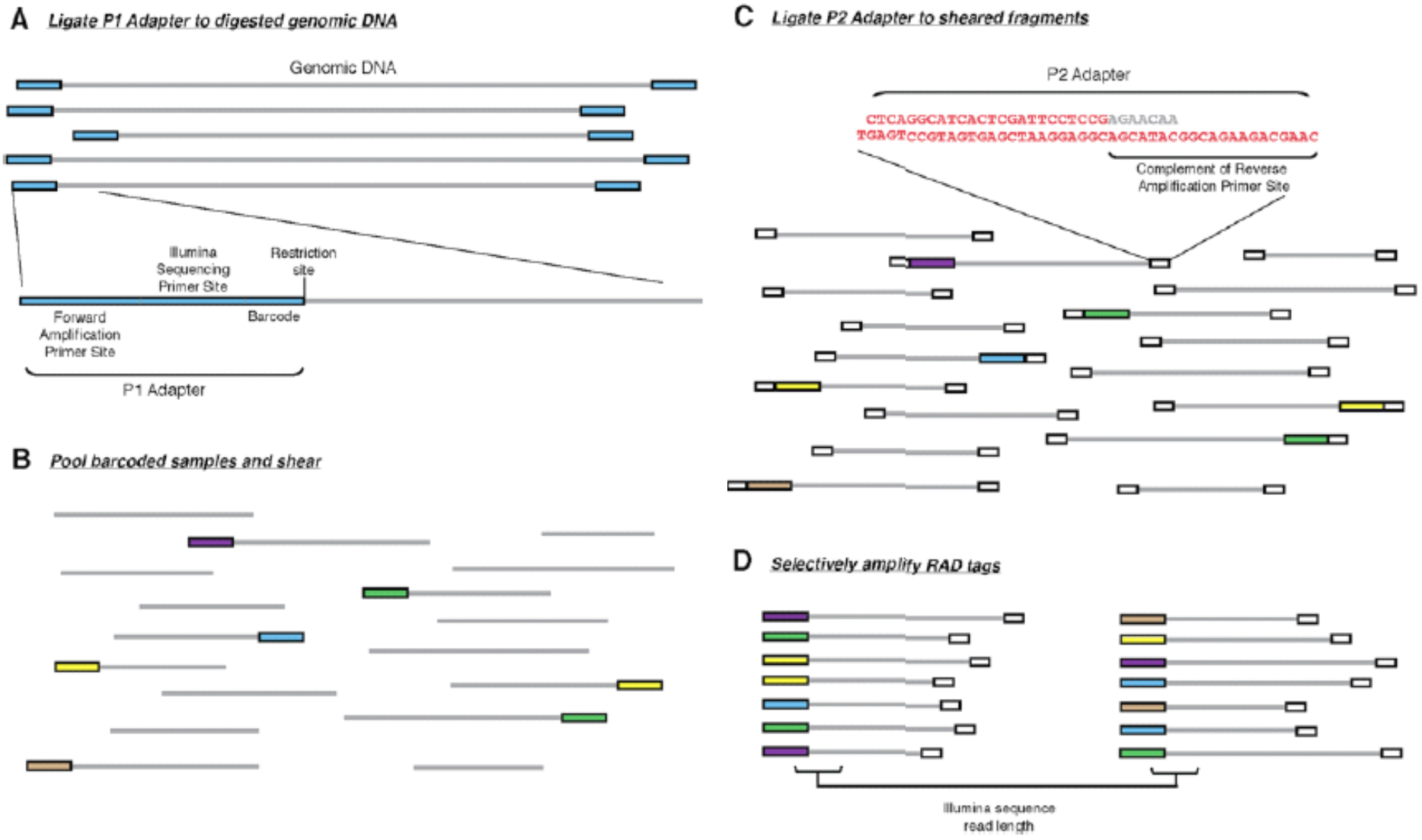
- Sample a fraction of the genome from multiple individuals and look for differences.
- Allows polymorphism discovery for genotyping and gives an idea of genome differences between hybrids.
- Pilot project (Linnean Society) digested standard Illumina paired end genomic libraries with MseI prior to final PCR step.
- Should remove 75% of genome. Compared *S. squalidus*, *S. vulgaris* and their two hybrids, *S. eboracensis* and *S. cambrensis*.
- Four libraries (one of each taxon) sequenced so far, data through quality control for two:

Average data return of 42 million reads per library.

RAD sequencing

- Provides a method for combining SNP discovery and genotyping into a single step.
- Genomic DNA is digested with a rare cutting restriction enzyme (selected to generate <300,000 fragments).
- Barcoded sequencing adaptors are ligated to the restriction site (allows pooling of multiple genotypes) prior to amplification.
- Because reads are adjacent to <300,000 possible restriction sites, sequence assembly is simple and coverage of each read will be high, allowing detection of SNPs within 50bp of the site.
- Also allows detection of restriction site presence/absence.

RAD sequencing



RAD sequencing

- Because # sites is low, a single next-gen sequencing run gives enough coverage to reliably detect polymorphisms between multiple pooled genotypes.

- For example: 300,000 fragments requires 3 million reads to achieve 10x coverage (minimum for SNP detection). Illumina sequencing returns ~15-30 million reads per lane, allowing 5-10 genotypes to be assayed (40-80 plants/run = small F2 population).

- Coverage needs to be higher if SNPs in the population are likely to differ between individuals, such as for genetic diversity studies:
  - Mapping population: sequence parents at 25x per allele
    sequence progeny at 10x per allele
  - Natural population: sequence all individuals at 25x/allele

RAD sequencing

- Genetic diversity studies in diploid *Senecio* species

- Using *S. squalidus* as example (1.7Gbp genome)

- SbfI digest would yield a predicted 30,000 RAD tags

So for 50x coverage (25x per allele) need 1.5 million reads

Thus can multiplex 10-20 samples in a single Illumina lane.

One lane costs ~£2000 to run.

Could genotype a separate population in each lane.

Alternatively, use RAD for SNP discovery and convert to an Illumina GoldenGate/Infinium assay for higher sample numbers.

Acknowledgements

University of Bristol:

Simon Hiscock (PI)

Keith Edwards (Co-I)

Gary Barker (Bioinformatics)

Jane Coghill (Transcriptomics)

Chris Thorogood & Sacha Allen
(Ph.D students)

Tom Batstone (Technician)

University of St. Andrews:

Richard Abbott (Co-I)

Adrian Brennan (Postdoc)

Aberystwyth University:

Wayne Powell (Director)

Charly Morgan (Technician)

Research Funding:

NERC
Linnean Society