Genome sequencing in *Senecio squalidus*

Outline of project

- A new NERC funded grant, the genomic basis of adaptation and species divergence in *Senecio* in collaboration with Richard Abbott and Dmitry Filatov
- In Bristol we are focusing on the molecular basis of changes to gene expression in *Senecio*, previously identified in our lab by Matt Hegarty
- Particular focus will be on miRNA profiling and the use of ChIP-seq to identify both changes in promoter sequence and activity
- These approaches require a reference genome sequence, and it is the production of this which we are currently working on

Partial reference genome of *S. squalidus* OX6

- Next generation sequencing technology is extremely powerful, but larger and more complex genomes still present a capacity challenge for *de novo* assembly
- Use of 2 platforms. 454 sequencing to produce longer, more accurate reads and Illumina sequencing for greater yield
- Focus on the non-repetitive regions of the genome to maximise the yield of coding sequence and facilitate assembly
- Present funding will produce only a partial genome, but we intend to supplement this work with the longer term aim of producing a complete genome sequence

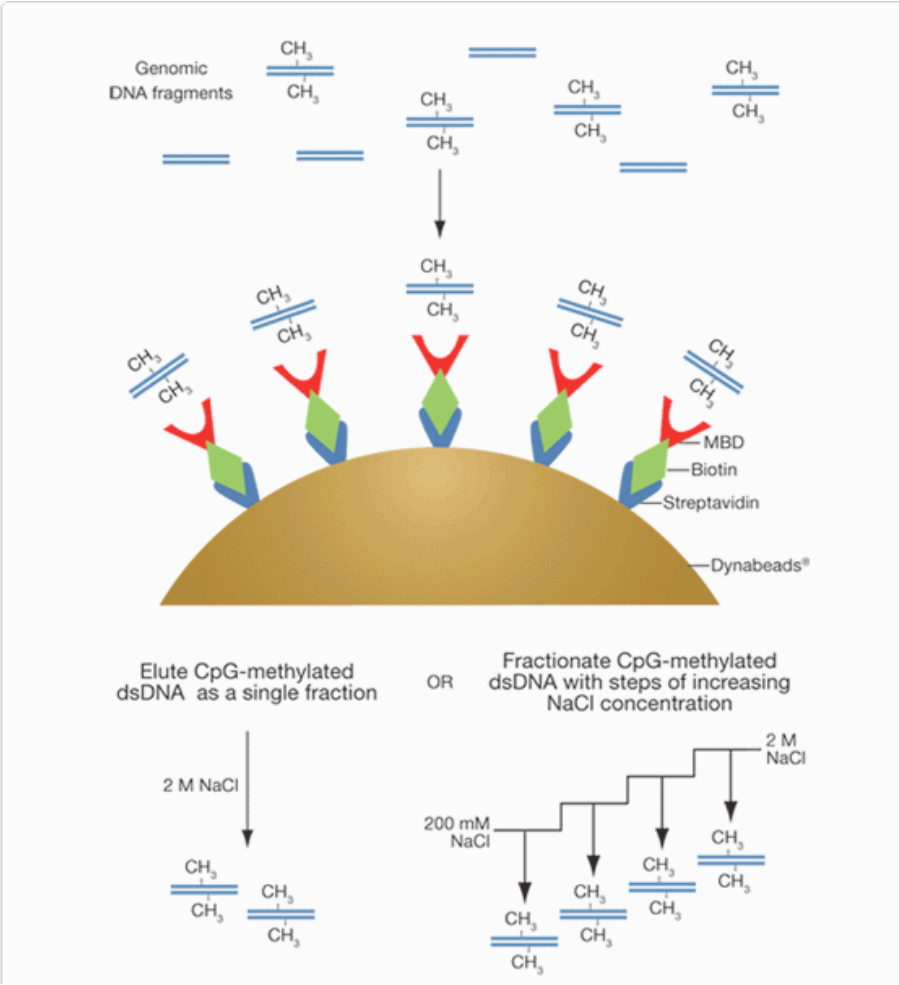
Sequencing strategy and approaches

- 454 sequencing of methyl-filtered libraries to produce a low coverage reference assembly
- Illumina sequencing of paired-end and mate-pair libraries to increase both genome coverage and sequencing depth
- Cot normalized Illumina libraries to complement the methyl-filtered data

Methyl-filtration

Repeat sequences in plant genomes have a tendancy to be heavily methylated, hence removal of the methylated fraction reduces their representation (Rabinowicz *et al.* 1999)

- Invitrogen MethylMiner kit
- Designed to purify sequences with a high proportion of CpG methylated sites
- Uses a methyl binding domain protein bound to magnetic beads to pull down target sequences
- Fractionation by variation of salt concentration during elution
- Filtered DNA used for preparation of libraries for 454 sequencing



Pilot 454 data overview

- Filtered and untreated libraries run on 1/16th of a 454 plate

	Methyl filtered	Untreated	Total
number of reads	107,138	107,842	214,980
Total sequence (Mbp)	32.97	39.84	72.81
Mean read length (bp)	311	379	-
Mode read length (bp)	412	500	-

Filtered vs untreated

- Filtered reads had approximately 10% more unique hits on SenecioDB
- Reads from each library were aligned using CD-Hit.
- Representative sequences from clusters comprising 30 or more sequences were Blasted against the Green Plants dataset at NCBI
- Of these larger clusters, all from the untreated library consisted of repeat elements, while all from the filtered data consisted of ribosomal sequence

Illumina sequencing

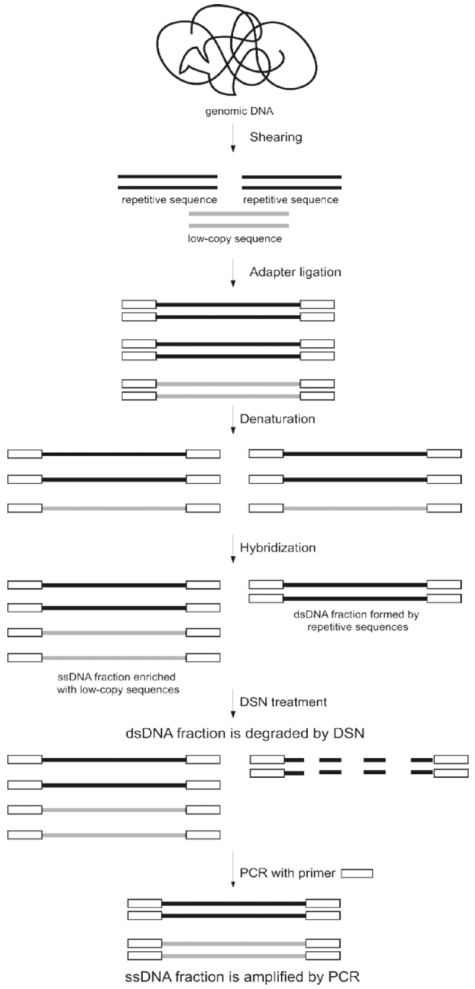
- Cost per base is ~40 fold less than 454 sequencing
- Use of paired-end and mate-pair libraries spanning a range of insert sizes will facilitate assembly
- 150bp read length is coming online now. We will pilot a run with this in the next couple of months
- Library preparation is carried out in-house, and sequencing is done by the transcriptomics facility in Bristol, so we have a greater degree of control over the process.

Illumina sequencing

- We have several libraries prepared with insert sizes ranging from 100 - 500bp
- A pilot run on a single lane using a 150bp insert paired end library with 76bp reads has yielded ~3.9Gbp of sequence that has passed quality filtering.
- This corresponds to ~2.3 fold genome coverage.

C_0t normalization of illumina libraries

- Method for enriching low copy number sequences
- Use of a thermostable DSN from crab rather than hydroxyapatite chromatography
- Used for normalization of plasmid and cDNA libraries
- Technique is easily adapted for normalization of illumina gDNA libraries



Shagina *et al.* (2010) Biotechniques **48**:455-459

Future genome development work

- 2 full 454 plates have been run at the NBAF Liverpool, and are currently in the data analysis pipeline.
- 8 remaining plates will be run over the next 2 months
- Optimisation of Cot filtration on illumina libraries and sequencing of these in Bristol.
- Sequencing of multiple insert size illumina libraries at the NBAF in Edinburgh
- Assembly and Annotation!

Acknowledgements

University of Aberystwyth
Matt Hegarty

University of Bristol
Simon Hiscock
Jane Coghill
Alexandra Allen
Keith Edwards

University of St Andrews
Richard Abbott

University of Oxford
Dmitry Filatov

NBAF Liverpool
Margaret Hughes
Christiane Hertz-Fowler
Andrew Cossins

NATURAL ENVIRONMENT RESEARCH COUNCIL

University of BRISTOL