**Case for Support Part 2**

**Background:**

*Cardamine hirsuta* is a diploid, self-compatible *A. thaliana* relative, which the Tsiantis group developed into a powerful model for studies in comparative plant biology with BBSRC funding (Barkoulas et al., Nature Genetics 40, 113 – 41; Blein et al., Science 322; 1835-183 Hay and Tsiantis, Nat. Genet. 38, 942-7). Despite being relatively closely related the two species differ considerably in developmental and biochemical pathways of agricultural relevance: morphogenesis and growth of shoots, flowers, leaves and fruits, glucosinolate biosynthesis, response to plant hormones and cellular pattern of lignin deposition in fruits. Thus, comparative studies between *C. hirsuta* and *A. thaliana* provide opportunities to understand genome evolution and its impact on diversification of these key traits and on recombination. For these reasons *C. hirsuta* is high priority for genome characterization in the international B-MAP Brassicaceae Map Alignment Project. A large collection of developmental mutants and two sets or RILs have been generated in Oxford providing a valuable lasting resource for comparative studies with *A. thaliana* and other Brassicas. These reagents and a large amount of incompletely assembled *C. hirsuta* genome sequence and transcriptome were developed in the context of recently ended or current BBSRC projects (BB/D010977/1; BB/F012934/1) but no funding exists to develop them to a level suitable for full comparisons to *A. thaliana* and dissemination to the scientific community.

While *A. thaliana* biological resources are well curated there is still a great need for comparative bioinformatics and biological resources that will allow exploitation of *A. thaliana* tools and knowledge base for understanding biodiversity. The time is particularly ripe for such activity as a high quality Arabidopsis reference genome exists and a large community of Arabidopsis researchers are interested in investigating biodiversity. Despite a lot of interest in natural variation, well-curated community resources that provide a roadmap to bridge the genotype to phenotype gap are still in their infancy. Additionally, the Arabidopsis community still lags behind that of other complex eukaryotes in terms of exploiting resources form model organisms to their relatives to understand biodiversity, for example the Drosophila 12 genomes project. Finally, despite isolated case studies there is a paucity of experimental platforms amongst *A. thaliana* relatives that show genetic tractability that is suitable for the analysis of gene function and evolution in a genome-wide and unbiased fashion. Our proposal helps to address these issues by providing a two-pronged resource that will allow the community to use *A. thaliana* as a springboard to identify and compare the genetic basis for variation in plant form and function within and between species. Our emphasis is on two types of resources: MAGIC lines designed to study *A. thaliana* natural variation, and *C. hirsuta* genome, RILs and mutants which will allow interrogation of the genetic basis for variation between species. Our rationale for selecting and integrating these resources is that while MAGIC lines are the richest resource available in terms of allelic diversity for understanding natural variation within a species in a crossing framework, the informatics support for their exploitation is limited. Conversely, *C. hirsuta* is an outstanding system to investigate variation between species as no other *A. thaliana* relative has a comparable level of experimental tractability, however community resources to support the broad use of this system are virtually non-existent. A distinguishing feature of the resource we proposes to generate is the direct integration of bioinformatics resources (genome sequences, transcriptome information, QTL data) and biological resources (RILs, MAGICs and mutants) for the purposes of comparative biology.

**Project aims and objectives:**

We aim to generate a complete set of (i) experimental reagents, (ii) genetic and genomic data, and (iii) bioinformatics tools and resources, for the integrated analysis of the plants *Arabidopsis thaliana* and *Cardamine hirsuta,* with emphasis on their evolutionary relationships and on the use of panels of recombinant inbred lines (RILs) for gene mapping and analysis of recombination processes.

We will pursue the following **specific objectives** in *C. hirsuta*:

1. Assemble a reference genome for *C. hirsuta*
2. Fully genotype three RIL sets and founders by re-sequencing
3. Generate a detailed synteny map between *C. hirsuta* and *A. thaliana* genomes
4. Map a panel of QTL and mutants for comparative analysis

**Applicants**

Professor Richard Mott is a principal investogator at the Wellcome Trust Centre for Human Genetics (WTCHG), University of Oxford. Over the past 20 years he has published in many areas of bioinformatics but for the past 5 years his research has focussed on the development and application of statistical methods and databases to dissect the genetic basis of complex traits in humans, mice and plants. He is responsible for the sequencing aspects of this project, for the databases and web browsers, and jointly responsible for the statistical analysis. The WTCHG is one the worlds leading genetics research laboratories with a strong emphasis on the development of statistical methods. The WTCHG Genomics and Bioinformatics Cores have strong expertise in genotyping and next generation sequencing.

**Programme and Methodology**

**1. *C. hirsuta* genome assembly**

**Background work.** An inbred reference strain of *C. hirsuta*, created by seven generations of selfing through single seed descent, has been used for genomic DNA and RNA sequencing with Roche 454 and Illumina platforms, cDNA and genomic BAC library construction, and generation of a genetic map*.*For this project we will provide Illumina genomic sequence at approximately 100x depth - half in the form of 51 bp reads from 400 bp insert libraries, and half in 76 bp reads from 3 kb insert mate-pair libraries. In addition, we will provide 3-4x depth of 454 genomic sequence of which a third is from 3 kb paired-end libraries. Currently 13.7 Gbp of Illumina sequence together with 0.8GB of 454 Titanium has been assembled in 91,883 contigs containing 154.4 Mbp. To aid assembly we will also provide 14,000 dideoxy-sequenced BAC ends. We have previously analysed these sequences to identify paired BAC end sequences that align with syntenic *A. thaliana* regions allowing us to anchor the *C. hirsuta* genetic map to the *A. thaliana* genome.

* 1. ***De novo* assembly algorithms**

Test algorithms designed for *A. thaliana* in new species Xxxx Richard to write a few lines.

We will work with TGAC in Norwich who have invited us to submit a CCC project to merge existing 454 and Illumina *C. hirsuta* data to help assembly of more complex crucifer genomes they are analyzing.

* 1. **Genome annotation**

Xxxx Richard to write a few lines.

**2. RIL genotyping**

**Background work.** Key to both QTL mapping and community exploitation of RILs is effective genotyping. To this end we will adopt a strategy of low patch sequencing that has been very effective in mice. Xxxx Richard to write a few lines.

**2.1 Light re-sequencing of RILs and founders**

We propose to use Illumina HiSeq2000 sequencing to generate 20x?? depth of genomic sequence for the founder lines of two RIL sets where the reference OX strain was used in common. We will follow similar methodologies that we used to assemble the MAGIC founder genomes in order to align the two new *C. hirsuta* genomes with the reference.

The genotyping strategy that we propose to follow for the three RIL sets, each containing 200 individuals, is a novel by-product of the RNA-seq transcriptome analysis that we have piloted on the Arabidopsis MAGIC RILs. By sequencing the transcriptome of each line, and aligning the reads to the reference genome, we can call sequence variants in exons as well as quantifying gene expression. The genome of each RIL is a mosaic of the haplotypes in the founders, so after filtering out uncertain variant calls (ie in repetitive regions or where mapping quality or PHRED quality is low), the remaining sequence variants will follow the mosaic structure of each line, and will permit us to map the recombination breakpoints in the genome mosaic to very high precision – typically under 10kb in data from a pilot RNAseq experiment our collaborator Richard Clark has performed in which the transcriptomes of seedlings from 30 MAGIC lines were sequenced. We (Mott group) developed a dynamic programming algorithm to deduce the most probable haplotype mosaic reconstruction of each line. We found we could reconstruct each genome with high accuracy – each unrecombined segment that matches to a single founder haplotype contains thousands of exonic polymorphic sites for which there is high quality read data. The high density of variable sites means that recombinations are mapped to about 10kb or less. Table 1 shows an example reconstruction for two chromosomes for the MAGIC line 12. The average error rate (ie sites where the allele called is inconsistent with the surrounding haplotype) is of the order of 10-3-10-4 discrepancies per site, equivalent to 10-6-10-7 discrepancies per bp of reference genome sequence.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **MAGIC** | **founder** | **chr** | **from bp** | **to bp** | **len/bp** | **len/sites** | **errors** | **errors/site** | **errors/bp** |
| MAGIC\_12 | mt-0 | 1 | 7202 | 739731 | 732529 | 784 | 1 | 1.28E-03 | 1.37E-06 |
| MAGIC\_12 | wu-0 | 1 | 739731 | 18333842 | 17594111 | 14386 | 14 | 9.73E-04 | 7.96E-07 |
| MAGIC\_12 | no-0 | 1 | 18333842 | 20036716 | 1702874 | 1617 | 0 | 0.00E+00 | 0.00E+00 |
| MAGIC\_12 | can-0 | 1 | 20036716 | 21142282 | 1105566 | 851 | 0 | 0.00E+00 | 0.00E+00 |
| MAGIC\_12 | no-0 | 1 | 21142282 | 28306201 | 7163919 | 8413 | 20 | 2.38E-03 | 2.79E-06 |
| MAGIC\_12 | hi-0 | 1 | 28306201 | 30416501 | 2110300 | 2200 | 1 | 4.55E-04 | 4.74E-07 |
| MAGIC\_12 | can-0 | 2 | 73391 | 6225662 | 6152271 | 2700 | 3 | 1.11E-03 | 4.88E-07 |
| MAGIC\_12 | zu-0 | 2 | 6225662 | 11457217 | 5231555 | 4902 | 5 | 1.02E-03 | 9.56E-07 |
| MAGIC\_12 | hi-0 | 2 | 11457217 | 12265326 | 808109 | 935 | 1 | 1.07E-03 | 1.24E-06 |
| MAGIC\_12 | no-0 | 2 | 12265326 | 19696711 | 7431385 | 6798 | 1 | 1.47E-04 | 1.35E-07 |

**2.2 Analyse cross-over distributions**

The high-resolution haplotype mosaics will enable us to map recombination hotspots and coldspots in both Arabidopsis and Cardamine RILs at very high precision. We will augment this data by also mapping ancient recombinants that occurred in the founder strains prior to the foundation of the RILs using the ancestral recombination graph analysis. This data will be used to characterize recombinants in both species, for example to search for motifs associated with hotspots. MORE

**2.3 RIL transcriptome analysis**

Following the methodologies detailed in xxx we will determine the transcriptome of seedlings from each RIL in each set, and of a subset of 40 RILs at three developmental stages from each set. The seedling RNAseq data will be used to determine haplotype mosaics as described above. We will sequence barcoded pools of 16? transcriptomes in parallel in a single lane of a hi’Seq in order to obtain a sufficient yield of sequence from each RIL at minimum cost. Cluster analysis will be used to identify global gene expression modules in *C. hirsuta* RILs. Statistical methods will also be developed to associate genomic regions with these expression modules in order to determine whether gene expression is controlled in *cis* or *trans* to each module. Comparative analysis of this data from both *C. hirsuta* RILs and *A. thaliana* MAGICs will allow us to (i) identify variable versus invariant components of gene regulatory networks (GRN) and (ii) pinpoint genomic regions that may harbor master regulators of GRNs in each species and determine to what degree they are conserved. In summary, these experiments will help ascertain to what degree conservation versus divergence of GRNs may contribute to diversity between *C. hirsuta* and *A. thaliana*. Patterns of natural variation in GRNs within each species will also provide a useful genomic context for future work to understand to what degree standing genetic variation in different species constrains natural variation and how this differs between closely related taxa (what is known from mice?). We will also map expression QTLs (both cis and trans) and identify loci that act as master regulators of modules by mapping QTLs controlling eigen genes for each module (ref)

We will compare the gene co-expression modules in the two species to identify conserved modules using the WGCNA package (ref Langfelder and Horvath 2008 BMC systems). We will search for over-represented sequence motifs in the neighbourhoods of the genes in each module and match these to known transcription-factor binding sites. For each module we will construct an eigen gene (a virtual gene whose expression levels are given by the first principal component of the expression data for the genes in the module) and map trans expression QTLs for it. This will help identify transcription factors that are likely to control the module,

In this way we can determine the degree of conservation of both gene regulatory networks and their molecular mechanisms across tissues and species. We will use this data to annotate the Arabidopsis and Cardamine genomes and make the data publicly available.

**3. Synteny browser**

**Background work.** To facilitate effective viewing and interrogation of the *C. hirsuta* and *A. thaliana* genomes in parallel, we will develop comparative databases and web interfaces. We have previously exploited to great effect the synteny between these two species for genetic mapping in *C. hirsuta*. A scaffold of paired BAC end sequences (BES), where both ends of a BAC reside within 250 Kb and are co-linear with orthologous *A. thaliana* sequences, were aligned across the *A. thaliana* genome to identify syntenic regions with considerable conservation of most genes lying between such paired BES. To expedite mapping we identified a 25K panel of high quality SNPs by mapping short reads from mRNA-seq of four polymorphic accessions to our reference transcriptome, and aligned these to the *A. thaliana* genome for convenient display on a Gbrowse database. An example of the utility of this database is shown in Fig. 1 where a *C. hirsuta* mutant for which no *A. thaliana* candidate gene could be identified was mapped to a 25 Kb region. This database has been a valuable in-house resource for mapping projects in the Tsiantis and Hay groups and provides proof of principle for a synteny brower with extensive functionality that we propose to develop here.

**3.1 Develop visualisation and interrogation tools for comparative genomics**

The annotated reference genome sequences for *A. thaliana* and *C. hirsuta* will be augmented by 19 founder genomes of *A. thaliana* MAGICs and 2 founder genomes of *C. hirsuta* RILs. Features for comparative genome interrogation will include the identification of conserved non-coding sequences and the integration of genetic map data with the physical genome.

**3.2 Develop statistical tools to apply genome data to RILs**

We will develop statistical tools to apply genome data to the *A. thaliana* MAGIC and two *Cardamine* RIL sets; e.g. imputation of the RIL genome sequences as mosaics of the founders, annotation of the effects of variants on gene structure, and superimposition of RNAseq data. We will develop tools to visualise differences between genomes and their consequences on transcription. Importantly, these computational tools will be applicable to other genomes.

**3.3 Compute ancestral recombination graphs**

We will combine genome sequence from 19 *A. thaliana* and three *C. hirsuta* genomes to compute ancestral recombination graphs across *A. thaliana*, using *C. hirsuta* as an outgroup. This data will tell us how the ancestral relationships within each species change across the genome, which each species acting as an outgroup to the other. We will use the BEAGLE and KWARG algorithms developed by Hein and Lynsgoe (refs). We will map recombinants and hot spots in the RI lines in both species (Henderson) to improve gene-mapping methods and to understand recombination processes better, e.g. by finding motifs associated with recombination. Cross-over frequency mapping information from *A. thaliana* will be used to test conservation of hotspot location in *C. hirsuta*.

In summary we propose to produce very deep functional and sequence annotations of the two species in order to understand better the consequences of their evolutionary divergence

**4. Map QTL and mutants**

**Background work.** To facilitate community use of the genetic resources developed here for *C. hirsuta* we propose to provide a first pass mapping of mutants and QTL for a panel of comparative traits. We have previously used QTL analysis to understand the genetic architecture of the dissected *C. hirsuta* leaf and to describe the loci affecting leaf shape and leaflet number. Three replicates of 195 genotyped RILs were phenotyped for leaf traits and composite interval mapping detected major QTL for all traits (Fig. 2). A second RIL set of 300 lines has been generated where the reference strain is a common founder and we propose to deposit both of these RIL sets with NASC.

**4.1 Map QTL for a panel of comparative traits in two genotyped RIL sets**

We will map QTL for the following traits in two RIL sets: hypocotyl length, leaf area, branch number, plant height and flowering time. These traits have been mapped previously in *A. thaliana* populations and will provide a first indication of the degree to which natural variation in the same or different loci contribute to these morphological traits. We have selected these traits for their agronomic importance in brassica crops such that the identification of *C. hirsuta* QTL may also provide useful information for robust crop improvement strategies.

**4.2 Characterise and map 15 EMS mutants for comparative analysis**

The analysis of induced variation in *C. hirsuta*,coupled with comparative studies in *A. thaliana*, has led to considerable new insights regarding the genetic basis for morphological differences and how these evolve (Tsiantis papers). We have catalogued a large collection of ethylmethyl sulfonate (EMS) mutants and propose to expedite community use of this resource by curating up to thirty heritable mutants to be deposited with NASC. To curate each mutant we will perform backcrosses to remove the bulk of unlinked mutations before phenotyping and mutants will be selected for comparable traits to those phenotyped in the RIL sets. Finally, we will pick 15 mutants that we judge to be more interesting in terms of community exploitation and determine their map position before dissemination to NASC. Additionally, we will provide an EMS mutagenised population for community screening.

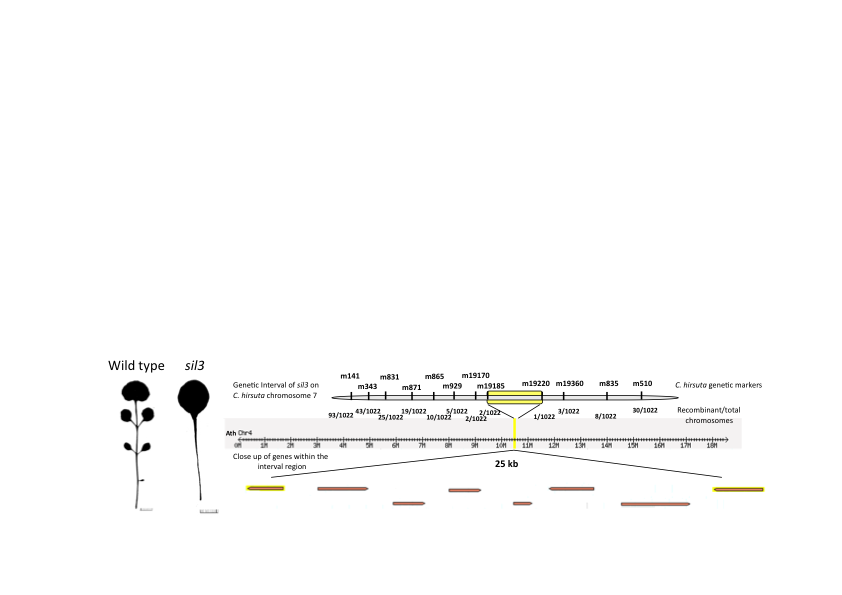


Figure 1.

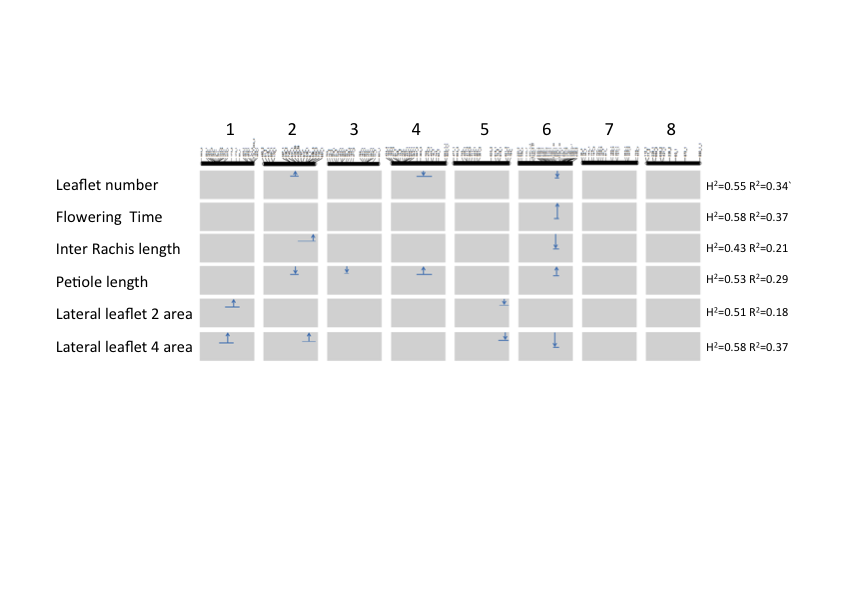


Figure 2.