Community Sequencing Program: **Project Proposal**

Proposer's Name: Rod A. Wing, Tom Mitchell-Olds, J. Chris Pires, M. Eric Schranz, Detlef Weigel, Stephen Wright

Project Title: Empowering functional plant genomics with genomes and transcriptomes of the Top 20 Brassicales

Proposal ID: 652

A) Brief description:

Abstract: The Brassicaceae constitute not only one of the most diverse plant families, but also one rich in agronomically important vegetable and oilseed crops. It has over 3700 species that grow in a wide range of environments and habitats. Several species have been domesticated: these include different cabbages, broccoli, turnip, rapeseed, horseradish, and several mustards. Brassicaceae include a number of species that are current and emerging biodiesel crops. *Arabidopsis thaliana*, arguably the world's most important and tractable plant experimental system, is also in this family. The strategy behind the current JGI proposal is to generate high quality genome and transcriptome data sets and associated analyses for the **Top 20 Brassicales** (18 Brassicaceae plus two outgroup species). The proposed assemblies, combined with presently available genome sequences as well as forthcoming sequences being generated by consortium members (several in collaboration with JGI) will allow us to perform a number of analyses, such as the identification of conserved noncoding sequences across the family, and genes and genomic regions subject to recurrent diversifying selection. More importantly, the proposed JGI data will empower our consortium members to launch numerous full genome sequencing projects aimed at finishing the **Top 20 Brassicales**, and beyond.

Scope of Work: We request that JGI produce low coverage draft genomic sequence and mixed stage transcriptome data from 18 Brassicaceae species plus two outgroup species (the **Top 20 Brassicales**). The 20 species are listed here by their phylogenetic lineages:

Lineage I (6 species related to *Arabidopsis*)- *Descurainia sophiodes* (200 Mb), *Smelowskia porsildii* (240 Mb), *Lepidium campestre* (250 Mb), *Rorippa barbareifolia* (280 Mb), *Cardamine hirsuta* (549 Mb), *Physaria fendleri* (630 Mb)

Lineage II (7 species related to *Brassica*) - *Myagrum perfoliatum* (325 Mb), *Caulanthus amplexicaulis* (372 Mb), *Thlaspi arvense* (560 Mb), *Crambe hispanica* (580 Mb), *Cakile maritima* (730 Mb), *Eruca vesicaria* subsp. *sativa* (740 Mb), *Vella anremerica* (890 Mb)

Lineage III and basal lineages (5 species at the base of the family) - *Euclidium syriacum* (260 Mb), *Diptychocarpus strictus* (300 Mb), *Malcolmia maritima* (310 Mb), *Alyssum linifolium* (300 Mb), *Arabis alpina* (420 Mb);

Outgroup species (2 species between Brassicaceae and papaya) – *Cleome violacea* (525 Mb), *Batis maritima* (445 Mb).

Low coverage sequence would consist of 100X PE Illumina with 280 bp and 500 bp inserts, and 20X MP Illumina with 5 kb, 10 kb and 15k b inserts. Transcriptome data would consist of 1-2 Illumina HiSeq lanes per species (19-38 lanes of data). RNA would be collected from a mix of tissues and developmental stages, as well as from various conditions important to the JGI mission (drought, salt stress, disease, etc.).

Our consortium would perform all genome assemblies and transcriptome mappings. We would display all assemblies and mappings using the JGI Phytozome interface. In addition, we will continue to develop a new genome browser that incorporates comparative information on conserved noncoding and coding sequence and regions subject to diversifying selection using the gold standard Arabidopsis genome sequence as a base.

B) Background information

Technical Information: Specific information on species and genome size is given below. G+C content, polymorphism level, and repeat structure for these species is not known unless indicated otherwise. However, most Brassicaceae have relatively small genomes with most highly repetitive regions restricted to pericentromeric regions. Also, most species selected are self-compatible and inbred, and hence highly homozygous.

Available Resources: For each of the species below there is seed, herbarium vouchers, photo documentation, and instructions for plant growth. All listed species have been identified by leading taxonomic experts (see letters). The only species with genomic resources already in place are for *A. alpina, C. hirsuta,* and *Caulanthus amplexicaulis*.

Technical Challenges: We do not anticipate any technical difficulties given our experience in growing the plant materials, isolating DNA and RNA, and our collective bioinformatics expertise.

Starting Materials: We have high quality DNA and RNA in hand, some of which has already been sequenced by PIs Wing and Pires in shallow sequencing runs. We are currently growing additional plants in growth chambers to acquire additional materials to begin the project in a timely fashion.

The 20 species are listed here by their phylogenetic lineages with genome size (in Mb), chromosome number, taxaonomic experts/stakeholders (in addition to the PIs) indicated.

01 Smelowskia porsildii Alpine Smelowskia (240 Mb ; 2*n* = 12)
Lineage I - tribe Smelowskieae (1 genus, 25 species)
Taxonomic experts/stakeholders: I.A. Al-Shehbaz
Notes: This species has a small genome and is sister to the rest of Lineage I.

02 Lepidium campestre Pepperweed, peppercress (250 Mb ; 2*n* = 16)
Lineage I - tribe Lepidieae (6 genera, 245 species)
Taxonomic experts/stakeholders: I. Al-Shehbaz, J. Gaskin, K. Mummenhoff, N. van Dam
Notes: A related species, *L. meyenii* (*L. peruvianum*), is an ethnomedicinal plant.

03 Descurainia sophiodes Northern tansy mustard (200 Mb ; 2*n* = 14)
Lineage I - tribe Descurainieae (6 genera, 58 species)
Taxonomic expert/stakeholders: B.E. Goodson, I.A. Al-Shehbaz
Notes: Many species are noxious weeds, and some species are toxic to livestock.

04 Cardamine hirsuta Bittercress (549 Mb ; 2*n* = 16)
Lineage I - tribe Cardamineae (14 genera, 352 species)
Taxonomic experts/stakeholders: I.A. Al-Shehbaz, K. Marhold, J. Lihová, N. van Dam.
Notes: The genus includes both self-compatible and self-incompatible species and is becoming a well-studied model for plant-microbe interactions (J. Dangl letter).

05 Rorippa barbareifolia Yellow-cress (280 Mb ; 2*n* = 16)
Lineage I - tribe Cardamineae (14 genera, 352 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz, P. van Tienderen, N. van Dam.
Notes: The genus includes both self-compatible and self-incompatible species.

06 Physaria fendleri Fendler's bladderpod (630 Mb ; 2*n* = 12, 14, 24)
Lineage I - tribe Physarieae (7 genera, 133 species)

Taxonomic expert/stakeholders: A. Doust, D. Baum, S. O'Kane Jr.
Notes: *P. fendleri* seeds have hydroxy fatty acids used in high-quality lubricants. It is also a model for fruit shape and trichome architecture (see Doust and Baum letters).

07 Thlaspi arvense Field penny-cress (560 Mb ; 2*n* = 14)
Lineage II - tribe Thlaspideae (12 genera, 78 species)
Taxonomic expert/stakeholders: K. Mummenhoff, M. Koch, I. Al-Shehbaz.
Notes: This species is an emerging biodiesel crop (http://afrsweb.usda.gov/is/AR/archive/nov10/biodiesel1110.htm?pf=1). It is becoming a well-studied model for plant-microbe interactions (See J. Dangl letter). Beilstein et al. (2010) used this genus as a fossil calibration point in his study on divergence times.

08 Myagrum perfoliatum Bird's-eye cress (325 Mb ; 2*n* = 14)
Lineage II - tribe Isatideae (9 genera, 98 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz.
Notes: This species is an annual herb and is the sister tribe to rest of lineage II.

09 Caulanthus amplexicaulis Claspingleaf wild cabbage (372 Mb ; 2*n* = 28)
Lineage II - tribe Thelypodieae (31 genera, 193 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz, A. Pepper.
Notes: This species is a model for studying adaptation to different soil types, as some populations are adapted to serpentine. A. Pepper (letter) has developed genetic resources. It is also related to *Streptanthus drepenoides*, a model for ecological speciation studied by S. Strauss (UC Davis), and to *Streptanthus polygaloides*, a model for metal hyper-accumulation studied by R. Boyd (Auburn).

10 Crambe hispanica Abyssinian kale (580 Mb ; 2*n* = 30)
Lineage II - tribe Brassiceae (47 genera, 235 species), subtribe Crambe
Taxonomic expert/stakeholders: S.I. Warwick, J. Hall (see Hall letter).
Notes: *C. hispanica* is grown for oil from the seeds that has similar characteristics to whale oil. Economic uses include ornamentals, vegetables, but particularly lipids (in lubricants, surfactants, plasticizers, paints, hydraulic & dielectric fluids).

11 Cakile maritima European Sea rocket (730 Mb ; 2*n* = 18)
Lineage II - tribe Brassiceae (47 genera, 235 species), subtribe Cakile
Taxonomic expert/stakeholders: J. Rodman, K. Donohue, J. Hall (see letter).
Notes: This species is a common exotic weed, particularly the coastal dunes of California.

12 Eruca vesicaria subsp. sativa Rocket, eruca, arugala (740 Mb ; 2*n* = 22)
Lineage II - tribe Brassiceae (47 genera, 235 species), subtribe Oleracea
Taxonomic expert/stakeholders: S.I. Warwick.
Notes: This species is grown in Europe and N. America as a salad plant, and in Asia for cooking oil and as food for animals. The oil is also used as an industrial lubricant and for cosmetic and medicinal purposes. The plant is also used as fodder for domestic animals.

13 Vella anremerica Vella (890 Mb ; 2*n* = 14)
Lineage II - tribe Brassiceae, (47 genera, 235 species) subtribe Vella
Taxonomic expert/stakeholders: S.I. Warwick.
Notes: *Vella* L. includes shrubs with two-segmented fruits, and represents the sister subtribe to the other lineages in the tribe Brassiceae, which includes the crop Brassicas.

14 Euclidium syriacum Syrian mustard (260 Mb ; 2*n* = 14)

Lineage III - tribe Euclideae (26 genera, 120 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz.
Notes: *E. syriacum* are white-flowered annual herbs native to Eurasia and are weedy in N. America. This species is one of the few small-genome sized taxa in Lineage III.

15 Diptychocarpus strictus (300 Mb ; 2*n* = 14)
Lineage III - tribe Chorisporeae (5 genera, 57 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz.
Notes: D. strictus is an annual in the cold desert of northwest China that produces heteromorphic fruits. This is one of the few small-genome sized taxa in Lineage III.

16 Malcolmia maritima Virginia stock (310 Mb ; 2*n* = 14)
Lineage III - tribe Anastaticeae (13 genera, 93 species)
Taxonomic expert/stakeholders: I.A. Al-Shehbaz, S. Warwick.
Notes: Virginia stock is a popular annual garden plant. It is native in the Mediterranean region where it occurs in maritime sandy habitats and in waste places.

17 Alyssum linifolium Madwort (300 Mb ; 2*n* = 16)
Unplaced Basal Lineage - tribe Alysseae
Taxonomic expert/stakeholders: I.A. Al-Shehbaz.
Notes: This species has one of the smallest genomes among the basal lineages.

18 Arabis alpina Alpine Rockcress (420 Mb ; 2*n* = 16)
Unplaced Basal Lineage - tribe Arabideae (15 genera, 520 species)
Taxon expert/stakeholders: M. Koch, K. Mummenhoff; I.A. Al-Shehbaz, G. Coupland
Notes: *A. alpina* is being developed into a model organism for population genetics.

19 Cleome violacea (Cleomaceae) Spiderflower (525 Mb ; 2*n* = 20)
Outgroup – sister lineage to Brassicaceae (Cleomaceae 17 genera, 150 species).
Taxonomic expert/stakeholders: J. Hall (see letter, developed VIGs for this species).
Notes: Critically important genome needed as outgroup to studies of ancient genome duplications and of evolution of C_4 photosynthesis in Cleomaceae.

20 *Batis maritima* (Bataceae) Turtleweed, pickleweed, saltwort 445 Mb 2*n* = 22
Phylogenetic outgroup – (Bataceae 1 genus, 2 species)
Taxonomic expert/stakeholders: R.F. Thorne, P. Edger.
Notes: Bataceae are halophytic shrubs with salty leaves and stems that have been eaten in salads or used medicinally to treat ulcers, kidney stones, and skin problems. *B. maritima* is native to N. American coastal salt marshes, the Caribbean, Hawaii, and S. America.

Project Description:

I. Importance of the research

The long-term goal of our consortium is to develop and exploit a within-family comparative genomics research platform that can be used to ask fundamental questions in plant evolutionary biology and to provide essential baseline knowledge for addressing the food and bio-energy needs of our planet for generations. The Brassicaceae family is the obvious choice to develop such a within-family comparative genomics system [1, 2] as it: 1) is an active target for the development of biodiesel crops; 2) is highly diverse, comprising over 3700 species that grow in a wide range of habitats; 3) includes many important vegetable and oil seed crops; 4) Brassicaceae species have small genomes (mean 1C-value of only 0.63 pg [3]) and 5) contains THE premier plant model system – *Arabidopsis.*

Our long-term vision is to generate ~150 reference genome sequences (not only re-sequences or gene-space assemblies) across the family, that can be used as a foundational research base upon which layers and layers of additional data can be added, e.g. transcriptomes, proteomes, epigenomes, metabolomes etc., along with genetic information. We feel that having access to high quality reference genomes (RefSeqs) is critical if we are to truly understand the biology of this important and keystone plant family.

Advantages for creating a within-family comparative genomics research platform are many- fold. First and foremost is that all questions can be addressed in a phylogenetically informed context (e.g. phylogenomics). Secondly, evolutionary process can be explored at whole genome levels – i.e. genome structure and content. Third, patterns of conservation and divergence can be elucidated, which will lead to a better understanding of selective constraints and targets of positive selection on genes and genomes. And lastly, such a rich research platform will aid in the development of more complex evolutionary models to further our understanding of genome-scale evolutionary processes.

To initiate our long-term goals and vision we propose to generate a data set that will allow us to establish a set of gene space assemblies and transcriptomes from our "top" 18 prioritized Brassicaceae species and 2 outgroup species. These data sets will provide a wealth of ground-breaking information from species that have little genomic data, but which are positioned at key nodes across the phylogenetic tree. In addition, the data sets and analyses will provide preliminary information that can be used by our consortium to apply for funding to transform our gene space assemblies into full fledge RefSeqs.

The generation of a within-family comparative genomics platform, as planned, would stand to be the most powerful such system for all plants and possibly all higher eukaryotes, and would be used to address some of most important and urgent problems the human race faces. In addition, these proof-of-concept experiments in the Brassicaceae can demonstrate feasibility for similar efforts in the economically important, complex genomes of the grass, sunflower and legume families. The fact that the majority of the functional biology could also be performed in Arabidopsis is also a clear advantage over other plant and animal systems.

II. Scientific questions you expect to answer

Our consortium, the Brassicaceae Map Alignment Project (BMAP) held 6 meetings over the past 2 years (http://www.brassica.info/resource/sequencing/bmap.php) and developed the following research themes. Such themes would be addressed using the Top 20 Brassicales data sets.

A. How much noncoding DNA is subject to selective constraints, and what is the functional importance of noncoding regions?

In comparison with other systems, our understanding of noncoding DNA function in plants has been relatively limited, and represents an important barrier for Arabidopsis functional genomics and crop improvement. We still know very little about what fraction of noncoding sequence is active, and how it functions in plant genomes. Comparative genomics provides an entry point into determining the roles of noncoding DNA; using sophisticated sequence comparisons [4], we can identify genomic regions under purifying selection, which can then be experimentally validated. This will also lead to the development of a new genome browser for the Arabidopsis community, potentially in association with TAIR and the International Arabidopsis Informatics Consortium, incorporating information on conserved noncoding sequence which will be a key and novel resource for functional genomics.

B. What genes/noncoding sequences have been the targets of diversifying selection, and what is the biological basis?

Comparative genomics also provides a key entry point into identifying the fraction of the genome under positive, diversifying selection, indicative of adaptive evolution. Genomic regions subject to current positive selection will show a recurrent substitution pattern where the rate exceeds reference neutral sites [5]. With the combination of these 20 genomes plus the available genomes generated by JGI and our consortium, we will have a powerful dataset for identifying such genes and noncoding regions. A phylogenetic approach to characterizing diversifying selection can only be successful with a significant number of genome sequences, to enable the characterization of recurrent rounds of positive selection. Integrating the patterns of constraint and diversification, we will be able to generate an unprecedented 'selection map' for the Brassicaceae genome.

C. What are the evolutionary consequences of whole genome duplication?

What factors determine the fate of duplicate genes? To what extent does polyploidy drive the relaxation of purifying selection or increased bursts of positive selection and diversification? How does genome evolution following polyploidization proceed over time? How often are hybridization and allopolyploidy triggers for adaptive radiation?

The Brassicaeacae show a remarkable number of recurrent Whole Genome Duplication (WGD) events, and contain many well-characterized ancient events of polyploidization [6]. This, combined with the detailed functional information from Arabidopsis, make for an ideal model system to better understand the role of WGDs in plant diversification and evolution. With the large-scale comparative sequencing in the

family, we will develop fantastic opportunities to examine the fate of genes following whole-genome duplication, and the processes of genome evolution that accompany polyploidization. For example, we can reveal whether specific genes or gene families are privileged targets for neo- or subfunctionalization of individual duplicates.

D. What factors govern genome and chromosome evolution?

A central question concerning genome evolution is the relative importance of mutation vs. natural selection. How much do changes in mutational properties vs. the efficacy and strength of selection drive genome evolution? With the introduction of a large-scale comparative framework, we will be able to examine in detail the role of these different processes in driving the evolution of genome size, base composition and genome structure. Brassicaceae genomes also show high levels of synteny arranged into conserved genomic blocks with genomic rearrangements occurring recurrently at particular regions [7]. Hence, with family-wide genomic data we can examine a number of breakpoints to address if there are fundamental mechanisms involved in genomic rearrangements across the family.

E. What are the genomic consequences of mating system evolution?

The evolution of selfing is one of the most common transitions in angiosperms, and is expected to have fundamental consequences for genetic diversity, the efficacy of natural selection, and the evolutionary fate of transposable elements [8]. The Brassicaceae family is characterized by a large number of repeated breakdowns of self-incompatibility, and a corresponding transition from outcrossing to selfing. With the advent of a large-scale comparative framework, we will be able to use replicate transitions from outcrossing to selfing to gain key genomic detail and statistical power for understanding the causes, consequences, and genetic basis of the evolution of selfing.

F. What is the basis of morphological variation?

In addition to selfing, a range of morphological and life-history traits has evolved repeatedly in the Brassicaceae, such as complex leaves, different fruit types, seed dispersal mechanisms and perenniality. Moreover, intra-specific variation in morphology has been exploited to select enormously diverse crops, such as kohlrabi, cabbage, Brussels sprouts, kale, cauliflower, all from the same species. Accordingly, some of the best work in the area of evolution of development (‘evo-devo’) has been done in the Brassicaceae [9]. With the Top 20 Brassicales genome sequences, this information can be greatly leveraged, by examining the genes shown to be responsible for intra- and inter-specific morphological variation. Of particular interest will be the intersection of these genes and their evolutionary signatures across the Brassicales with genes identified by the approaches discussed above as being under diversifying selection.

G. Hybrid vigor/heterosis

Perhaps the greatest unsolved problem in plant biology remains the molecular and genetic basis of positive heterosis or hybrid vigor [10]. Although the phenomenon that F_1 hybrids outperform their parents is widely exploited in plant breeding, the underlying causes are largely unknown, with the notable exception of a few cases, where individual genes that can explain certain aspects of heterosis have been identified. The Top 20

Brassicales will enable systematic studies to determine the contribution of genome-wide phenomena to hybrid vigor.

H. Chemical variation (glucosinolates, etc); plant/animal interactions.

The major chemical defenses in the Brassicales are the sulfur-containing glucosinolates [11]. Large differences in glucosinolate profiles exist. Much of the variation across the Brassicales is due to differences in amino acid precursors, with four major classes. Glucosinolates derived from Val, Ile, Leu, or Ala are aliphatic, those from Phe or Tyr are aromatic, and those from Trp are indole glucosinolates. The fourth class, the Met-derived glucosinolates, is unique sub-set of aliphatic glucosinolates to the most speciose family, the Brassicaceae. Most of the genes controlling the formation of glucosinolates have been worked out in Arabidopsis and Brassica species. The Top 20 Brassicales species listed in this proposal have a wide-spectrum of glucosinolate profiles. For example, *Cleome violaceae*, *Batis maritima Lepidium campestris*, *Rorippa islandica* and *Cardamine hirsuta*, all lack Met-derived glucinolates. Hence, the proposed data will provide a unique opportunity for comparative functional and evolutionary analysis of the key chemical defense pathway and provide opportunities for future modification of defenses in key crop species.

Additional opportunities include:

I. Comparative systems biology of Brassicaceae; rhizosphere/metagenomics; synthetic biology

J. Comparisons of annual vs. perennial at the whole genome level.

K. Australian crucifers experienced a whole genome duplication and extensive genome reshcuffling (diploidization). These species have an interesting repertoire of repetitive elements, heterochromatic arrays, and remarkable phylogeographic history.

L. Uncovering the basis of physiological adaptations (edaphic, stress) of "extremophile" species/"extreme" traits of heavy metal hypertolerance and hyperaccumulation.

III. Preliminary data produced in support of the project

As the first steps towards large-scale comparative genomic analysis in the Brassicales, our international consortium has initiated a study of eight Brassicaceae genomes in collaboration with JGI. This study includes four JGI-sequenced genomes (*Arabidopsis lyrata*, *Capsella rubella*, *Thellungiella halophila*, and *Boechera stricta*), as well as three genomes sequenced by a Genome Canada/Genome Quebec-funded project led by Thomas Bureau (McGill University) and Stephen Wright (University of Toronto).

Below, we briefly describe some of the highlights of the *A. lyrata* and *C. rubella* work, followed by a presentation of the work done by the Canadian consortium.

The* Arabidopsis lyrata *genome sequence [12]. The genome of the self-incompatible perennial *A. lyrata* is larger than 200 Mb, near the family average for the Brassicaceae,

while the self-compatible annual *A. thaliana* has one of the smallest angiosperm genomes, at about 125 Mb, even though the two species diverged only about 10 million years ago. A high-quality genome sequence for the partially inbred *A. lyrata* strain MN47 was assembled by JGI from approximately 8.3x coverage of dideoxy sequencing reads, making use of information from genetic maps and chromosome painting. The final assembly included 206.7 Mb of sequence. A combination of de novo predictions, homology to *A. thaliana* features, and RNA sequencing was used to annotate the genome. In *A. lyrata*, we predicted 32,670 protein-coding genes, compared to 27,025 genes in *A. thaliana*.

Apart from chromosomal-scale changes, approximately 90% of the two genomes have remained syntenic, with the great majority in highly conserved collinear arrangements. A whole-genome alignment revealed that more than 50% (<114 Mb) of the *A. lyrata* genome appears to be missing from the *A. thaliana* reference genome. In contrast, only about 25% (<30 Mb) of the *A. thaliana* genome is absent from *A. lyrata*. Nevertheless, the distribution across different sequence classes is similar: half of the unalignable sequences are in TEs, and a quarter in intergenic regions. The net effect of these changes is that the *A. thaliana* genome is ~80 Mb smaller than the *A. lyrata* genome, with a much higher fraction of genic sequences. The apparent shrinkage of the *A. thaliana* genome is not simply due to a few chromosome-scale changes: only 10% of the size difference is attributable to the three missing centromeres; the rest can be attributed to hundreds of thousands of smaller insertions and deletions, spanning all classes of sites. Strikingly, while large differences much more often correspond to sequences only found in *A. lyrata*, this is not true for very small insertions and deletions. This is in stark contrast to other genomes from other closely related species, but with similarly sized genomes, such as chimpanzee and human.

As in other taxa, TEs make an important contribution to the change in genome size, and TEs not only comprise a larger fraction of the *A. lyrata* genome, but are also differently distributed in the two species. *A. lyrata* has a substantially higher proportion of genes with a TE nearby than *A. thaliana*, and this distance is skewed towards larger values in *A. thaliana* Together, these observations are consistent with a model under which selection purges TEs with deleterious effects on adjacent genes, such that TEs more distant from genes preferentially survive, with TE elimination having been more efficient in *A. thaliana*.

We used within-species polymorphisms to shed light on the process by which the difference in genome size between *A. thaliana* and *A. lyrata* arose. The observed patterns suggested that deletions are favored over insertions because of selection, rather than simple mutational bias, thus leading to a smaller genome in *A. thaliana*.

In summary, even the comparison between the genomes of two species has already been enormously informative, and it is easy to see how additional genomes will further increase the knowledge that can be gained from this type of approach.

The Capsella rubella genome. The current reference (JGI pre-release vHTZF) of *C. rubella* is, at ~150Mb, somewhat shorter than the expected genome size of 215Mb. The difference is being investigated but may reflect a centromeric expansion since a high proportion of unmapped reads could be tentatively associated with centromeric sequences previously identified. To validate the intra-scaffold integrity as well as to infer the

location and ordering of scaffolds for assembly purposes, we generated a high SNP density linkage map using an interspecific (*C. rubella* x *C. grandiflora*) genetic mapping population of 550 F2's and an Illumina-sequencing approach. Our genotyping generated more than 70,000 SNPs and a linkage map based on the JGI pre-release assembly.

In general, the assembly data combined with our linkage mapping confirm early comparative mapping results that suggested strong synteny between *C. rubella* and *A. lyrata*. Additional data suggested that sites of discontinuity with *A. lyrata* likely reflect false joins in the genome assembly, rather than true rearrangements between species. Resolution of breakpoints in the assembly has been aided by our linkage mapping, by identifying the closest adjacent markers that show a drop in linkage.

Preliminary analyses suggest that the euchromatic portion of the genome of *C. rubella* is consistently smaller than in *A. lyrata*, highlighting that a three-way analysis with *A. thaliana* will be important for inferring the directionality of species changes. Analysis of SNP variation and divergence between *C. rubella* and its close outcrossing congener *C. grandiflora* reveal signs of a genome-wide relaxation of purifying selection in the selfing species, consistent with theoretical predictions on the effects of linkage and a reduction in effective population size. We have also identified candidate differentially expressed genes that may be associated with floral evolution in the transition to selfing.

***Initial family-wide comparisons*.** The Canadian consortium has produced Illumina-based assemblies for *Aethionema arabicum* (*Aethionema* is the sister group to the rest of the Brassicaceae [13]), *Sisymbrium irio* (closely related to *Brassica*, but lacking the *Brassica* whole-genome triplication event), and *Leavenworthia alabamica*. These latter three genomes are being sequenced using an Illumina sequencing strategy that is comparable to the one here for the **Top 20 Brassicales**. We therefore present in some detail our assembly approaches and first analysis of conserved noncoding sequence below. Our initial analyses have been conducted using six species, including the three sequenced by the Canadian Consortium, *Arabidopsis thaliana*, *Arabidopsis lyrata* (JGI), and a pre-release version of *Capsella rubella* (JGI).

For the initial sequencing of *Ae. arabicum*, *L. alabamica* and *S. irio*, each species utilized 2 to 3 lanes of Illumina Genome Analyzer II data of 2x108 bp paired-end (PE) sequences with 280 bp inserts; and 1 lane each of mate-pair (MP) sequencing of 2x108 bp with an insert size of 5 kb. This produced ca. 30 Gb of data for each species. The genome assembly process began in May 2010 and, after a series of successive improvements in assembly quality, by April 2011 had led to contig N_{50}s of approximately 30 kb, and assembled scaffolds ranging from N_{50} ~122 kb (*Leavenworthia*) to 235 kb (*Sisymbrium*). The assemblies were frozen in September 2010 for initial annotation and initial comparative genomic analysis. The contigs available at this stage comprised over 98% called bases. As repetitive regions appear less degenerate to a de Bruijn-type assembler as k-mer length increases, we anticipated that repeat content assessed from custom repeat annotation of contigs would rise with k-mer length; this was indeed observed (e.g., ~12% at k=31 to ~24% at k=51 in *L. alabamica*). In December 2010 we incorporated the MP data. The initial MP insert sizes had been informed by simulations conducted on repeat size distributions. This allowed for the scaffolding of contigs separated by single instances of most repeat classes (<8 kb), and assembly of initial scaffolds having N_{50} ~ 150 kb from the mixture of MP and PE data for the 3 genomes.

While this reduced called bases to ~95% of the larger scaffolds (>500 bp), it extended contig length and corrected some long-range ordering problems within contigs. The MP data are currently incorporated into the de Bruijn assembly phase. Strategies are being considered that will allow additional information from MP data to be more fully integrated in a modified Hamiltonian scaffolding strategy, which will also allow scaffolding of contigs separated by longer arrays of repetitive regions. We are currently adding 10 kb MP data.

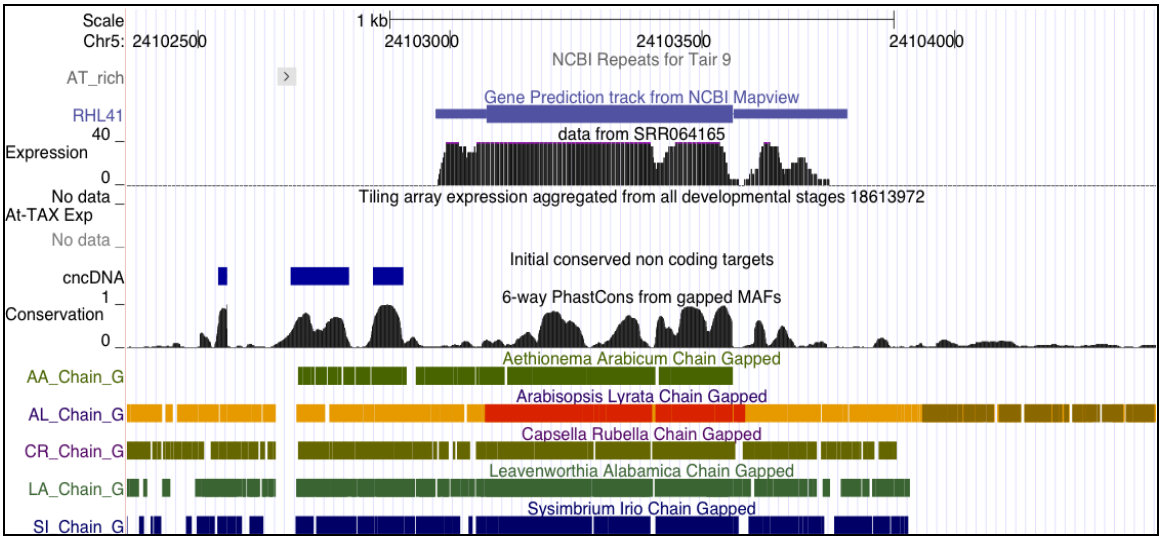
The total genome build sizes for *S. irio* and *Ae. arabicum* are currently ~80% of the expected total genome sizes, while the build of *L. alabamica* is only ~40% of the expected size of 450 Mb. In this species, preliminary surveys of repeat content indicate large repeat/transposon expansion while segregation patterns do not suggest polyploidy, but this is being further investigated. Efforts to finish these genomes are underway, including the use of BAC-end sequencing of *S. irio* in collaboration with Chris Town at JCVI, and the use of KeyGene BAC profiling data for *Ae. arabicum* generated by Eric Schranz. For all three species, we are generating F_2 mapping populations, and will use Illumina-based genotyping approaches to aid in large-scale scaffold ordering and identification of false joins. This approach has recently proved useful for aiding in the assembly of *Capsella rubella* through a genetic map generated by the Wright lab in an ongoing collaboration with JGI (above).

Multiple genome alignment. Our first comparative genome analysis was conducted using six species. The genome of *A. thaliana* was aligned pairwise with the genomes of five species: the fully assembled *A. lyrata* [14], the partially assembled pre-release genome of *C. rubella* (JGI) and the sets of contigs (September, 2010 annotation freeze) from *L. alabamica*, *S. irio* and *Ae. arabicum*. While the *A. thaliana* genome was repeat-masked at source (NCBI), the others were soft-masked using RepeatMasker [4] with a combination of *A. thaliana* repeat sequences and the species-specific repeat library generated as described above. Tandem repeats were left unmasked. We used LastZ [15] in both gapped and ungapped alignment modes for pairwise alignment. The local alignments were chained into longer sets of co-linear blocks using axtChain [16], then sorted and filtered for display as both gapped and ungapped alignment tracks on our genome browser. Separately the gapped alignments were mutually aligned using MultiZ [17], resulting in an *A. thaliana*-referenced 6-way multiple alignment. The alignments reveal large syntenic chains that frequently span whole contigs, highlighting that few genome rearrangements may have occurred in the Brassicaceae.

***Detection of conserved regions*.** Conservation scores were generated using PhastCons, which computes the posterior probability (0 to 1) that the region of multiple alignment is under evolutionary constraint [18]. As expected, the vast majority of *A. thaliana* annotated protein-coding exons show high conservation with the other five species. 59% of the Arabidopsis exons had extended regions of moderate to high conservation relative to the other species (10% or more of the exons had a PhastCons score of at least 0.5). For non-coding regions, PhastCons parameters were adjusted to increase sensitivity to shorter

conserved regions (expected length 20) at the cost of sometimes fragmenting large functional regions.

Importantly, we identified a set of 65,690 conserved regions that do not overlap annotated exons. The size of these regions ranges from 7 to 2460 bp, with an average of 52 bp. Many of these regions show a degree of conservation similar to average coding exons, yet show no evidence of expression or coding potential. To classify the non-coding conserved regions according to their location, we obtained the coordinates for the 5' promoter region of genes (-500nt), downstream region of genes (+500nt), intronic regions, intergenic regions, as well as regions noted as ambiguous . Gene annotation data was taken from TAIR9 (http://www.arabidopsis.org) and supplemented with RNA sequencing data aggregated from six experiments covering different tissues and stress conditions. We found a roughly equal number of conserved regions in each annotated class of locations (i.e. 5' promoter, 3' downstream, intronic and intergenic). However, because promoters and intronic regions account for a relatively small fraction of the complete genome, this translates a great enrichment in promoters for conserved non-coding DNA, with a coverage of 7%, whereas intronic and intergenic regions contain 2% and 5% conserved non-coding DNA respectively. To our knowledge this represents the first genome-scale identification of conserved non-coding regions in Arabidopsis, and confirms the large-scale presence of these sequences. We now have a first iteration of an Arabidopsis genome browser that incorporates information on conserved noncoding sequences across the six species (see sample figure). Thus, we have the analytical pipelines well in place to build a large-scale comparative genomics analysis through the sequencing of the Top 20 Brassicales.



Displayed is the neighborhood around *Arabidopsis thaliana* gene RHL41 along with tracks produced by the Canadian Consortium. The locus shows differential expression under cold conditions and was used as one of a small set of genes tested for conserved upstream motifs. The tracks include, from top to bottom: gene annotation from TAIR9

(http://www.arabidopsis.org); gene expression aggregated from tiling array data produced by the Weigel lab (At-TAX expression atlas); our tracks annotating conserved noncoding regions, PhastCons data on sequence conservation, and both gapped and ungapped alignment from our study species and *Arabidopsis lyrat*a.

IV. Size and nature of the larger community that will use the data

The BMAP consortium is composed of approximately 200 scientists that are highly connected to their individual crop and species-specific communities (http://www.brassica.info/resource/sequencing/bmap.php). The Brassica and Arabidopsis communities easily comprise 10,000 researchers.

These groups will benefit from our proposal research because the grand challenge questions to be addressed can now be asked within a phylogenetic framework which is extremely powerful in the detection of genes and genomic regions that are under various forms of evolutionary selection, such as those related to adaptation to extreme and low-nutrient environments, which is highly relevant to the bioenergy field.

V. Relevance of the project to the DOE mission and/or its economic or societal importance.

Numerous species in the Brassicaceae are current and emerging biodiesel crops, including canola and *Camelina sativa*. In addition to the oil-rich seed, the ability of Brassicaceae species to grow on marginal land with minimal inputs make them particularly attractive and potentially viable biofuel candidates, and this highlights the need to better understand the genomic basis of adaptation to extreme environments, as well as traits associated with biofuel production itself. To achieve this, a large-scale comparative framework is absolutely essential, in order to: 1) identify genomic regions that are key for adaptation to marginal habitats; 2) harness the power of *Arabidopsis* genetics and genomics, and 3) better understand functional noncoding DNA, which is likely to be key for future crop improvement based on genomic selection. With the large-scale comparative approach in tandem with population genomics approaches, one goal will be to create a 'selection map' that identifies key targets of selection associated with important agronomic traits.

In addition, one of our focal species is the most recent target for biodiesel production in the Brassicaceae, *Thlaspi arvense* (field pennycress). Recent research by the USDA shows promise for this common roadside weed as an important biodiesel crop in rotation with others such as soybean. The genomic resources for this species generated in this proposal will provide a key first step towards crop improvement for both further improvement of oil seed traits, as well as harnessing natural variation for genomic selection for key traits such as flowering time and nitrogen use efficiency.

In addition to the Brassicaceae species we propose to sequence, we have also included two outgroup species. Of particular relevance to the DOE is *Cleome* [19]. Currently, the Beijing Genome Institute (BGI) is sequencing two species, *C. gynandra* (a C_4 species) and *C. spinosa* (a C_3 species). In this proposal, we list *C. violacea* (a C_3

species), which is one of the most basal lineages within the family Cleomaceae and hence is critical for rooting evolutionary analyses of C_4 pathway evolution. Additionally, *C. spinosa* and *C. gynandra* share an ancient whole genome duplication, which is lacking from *C. violacea*. The Cleomaceae represents the most closely related C_4 plant lineage to Arabidopsis. Studying the evolution of C_4 photosynthesis in *Cleome* fits squarely with the key objectives of DOE to increase biomass of potential biofuel crops, to study carbon cycling and to improve genome phylogenetic coverage. The components and steps of dicot C_4 evolution identified by *Cleome* genomics can directly be applied to bioenergy research. Engineering C_4 has the potential to lead to higher yields and increased stress tolerance, issues of fundamental importance to bioenergy production and crop improvement.

The Top 20 Brassicales sequence data proposed here will also enhance the DOE mission to improve phylogenetic coverage. In particular, *Cleome* and *Batis* represent a bridge between the Brassicaceae genomes and the more distantly related papaya (Brassicales) and other Rosid genomes. This project will facilitate exploiting the wealth of data obtained from the premier plant research system, *Arabidopsis thaliana*, and the transfer of this information to more distantly related bioenergy and crop species.

D) References:

1. Franzke, A., et al., *Cabbage family affairs: the evolutionary history of Brassicaceae.* Trends in Plant Science. 16(2): p. 108-116.
2. Schranz, M.E., et al., *Comparative genomics in the Brassicaceae: a family-wide perspective.* Current Opinion in Plant Biology, 2007. 10(2): p. 168-175.
3. Lysak, M.A., et al., *The Dynamic Ups and Downs of Genome Size Evolution in Brassicaceae.* Molecular Biology and Evolution, 2009. 26(1): p. 85-98.
4. Stark, A., et al., *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.* Nature, 2007. 450(7167): p. 219-32.
5. Yang, Z., *Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.* Molecular biology and evolution, 1998. 15(5): p. 568-73.
6. Barker, M.S., H. Vogel, and M.E. Schranz, *Paleopolyploidy in the Brassicales: Analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales.* Genome Biology and Evolution, 2009. 1: p. 391–399.
7. Schranz, M.E., M.A. Lysak, and T. Mitchell-Olds, *The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes.* Trends in Plant Science, 2006. 11(11): p. 535-542.
8. Charlesworth, D. and S.I. Wright, *Breeding systems and genome evolution.* Current opinion in genetics & development, 2001. 11(6): p. 685-90.
9. Bowman, J.L., *Molecules and morphology: comparative developmental genetics of the Brassicaceae.* Plant Systematics and Evolution, 2006. 259(2-4): p. 199-215.
10. Chen, Z.J., *Molecular mechanisms of polyploidy and hybrid vigor.* Trends in Plant Science. 15(2): p. 57-71.
11. Windsor, A.J., et al., *Geographic and evolutionary diversification of glucosinolates among near relatives of Arabidopsis thaliana (Brassicaceae).* Phytochemistry, 2005. 66(11): p. 1321-1333.
12. Hu, T.T., et al., *The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.* Nature Genetics, 2011. 43(5): p. 476-+.
13. Al-Shehbaz, I.A., M.A. Beilstein, and E.A. Kellogg, *Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview.* Plant Systematics and Evolution, 2006. 259(2-4): p. 89-120.
14. Hu, T.T., et al., *The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.* Nature genetics, 2011. 43(5): p. 476-81.
15. Schwartz, S., et al., *Human-mouse alignments with BLASTZ.* Genome Res, 2003. 13(1): p. 103-7.
16. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. 12(6): p. 996-1006.
17. Blanchette, M., et al., *Aligning multiple genomic sequences with the threaded blockset aligner.* Genome Res, 2004. 14(4): p. 708-15.
18. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. 15(8): p. 1034-50.
19. Brown, N.J., K. Parsley, and J.M. Hibberd, *The futured C4 research - Maize, Flaveria or Cleome?* Trends in Plant Science, 2005. 10(5): p. 215-221.

E) CV's of lead investigators: 6

Tom Mitchell-Olds, Duke University

J. Chris Pires, University of Missouri

Eric Schranz, University of Amsterdam

Detlef Weigel, MPI for Developmental Biology

Rod A. Wing, University of Arizona

Stephen I. Wright, University of Toronto

F) Letters of support: 17

Ihsan AI-Shehbaz, Missouri Botanical Garden, USA

David Baum, University of Wisconsin, USA

Mark A. Beilstein, University of Arizona, USA

Mathieu Blanchette, McGill University, Canada

Jeff Dangl, University of North Carolina, USA

Andrew Doust, Oklahoma State University, USA

Jocelyn Hall, University of Alberta, USA

Marcus A. Koch, Heidelberg Botanic Garden & Herbarium, Germany

Gerhard Leubner, University of Freiburg, Germany

Martin A. Lysak, Masaryk University, Czech Republic

Klaus Mummenhoff, University of Osnabreuk, Germany

Alan Pepper, Texas A&M University, USA

Steve Robinson, Agriculture and Agri-Food Canada, Canada

Daniel Schoen, McGill University, Canada

T.F. Sharbel, Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Germany

Nicole M. van Dam, Radboun University, Netherlands

Peter van Tienderen, University of Amsterdam, Netherlands