Prospective project for the continuation of the international consortium for the development of genomics, genetics and bioinformatics resources for sunflower

Partnership

Public laboratories: University of British Columbia, University of Georgia, French National Institute of Agronomical Research (INRA)

Private partners: up to now Biogemma (representing Soltis and RAGT2n), Syngenta, Pioneer, Dow, Advanta, KWS, and Bayer. Potential new partners include MaisAdour, Caussade, and NuSeeds.

Context

The Sunflower Genomic Resources Consortium was created in 2012 in conjunction with a Genome Canada project led by L. Rieseberg's lab in collaboration with UGA and INRA. The main goal was to pursue the production of the first high quality genomic sequence of sunflower genotype HA412-HO.

In January 2015, the consortium partners decided to stop the assembly efforts and consider the "bronze" version of HA412-HO as the first reference sequence despite some assembly problems at the local scale. This reference sequence was made available to the consortium partners in January 2015 and 6 months later publicly at https://www.heliagene.org/HA412.v1.1.bronze.20141015/ .

Concomitantly, the emergence of PacBio technology enabled INRA to initiate the sequencing of XRQ and PSC8 lines using this technology in the framework of the SUNRISE and HELIOR projects.

In September 2015, initial results from PacBio sequencing of XRQ to 100X depth, indicate clearly that the sequence to be produced will be of much better quality. However, this genome will still be incomplete (about 20% of the sequence missing), and mis-assembly issues are likely due to scaffolding based on genetic markers only. Non-recombining regions and identical regions between parental lines of mapping populations will therefore likely contain local scaffolding errors. The production of pseudomolecules and annotation of the XRQ genome sequence is under way and should be made available to the consortium through the INRA website www.heliagene.org in early 2016.

In addition, the sequencing of the PSC8 line is under way with a final coverage of around 50X. It has started to produce results that we have analyzed at the candidate locus level. Although gene order is fairly well conserved across both genotypes, intergenic regions, comprising very high levels of repeat elements, were very different and could hardly be aligned. The release date of the PSC8 sequence has not yet been determined.

The consortium reached its goal to produce the first annotated genomic sequence of sunflower based on data produced by the Genome Canada project and INRA. It provided to the consortium members major resources such as BAC libraries ready to screen for fine-mapping, a physical map, reference transcriptomes with organ-specific expression patterns, a high quality and high density genetic map, and re-sequencing data for the 288 genotypes of the sunflower association mapping population. In addition, it secured the production of a second genomic sequence of higher quality in 2016. These resources will be a cornerstone for sunflower genetic and genomic research and breeding. Importantly, this consortium was a unique forum to bring together key international actors on one hand from the academic community and on the other hand from the biotech and seed companies.

However, our current analysis of structural variation from high quality PacBio assemblies clearly shows that our knowledge of sunflower genomes is far from satisfactory, which will impede fast and efficient QTL fine-mapping and the introgression of advantageous alleles. We therefore believe that it is necessary for our community to continue these efforts and further develop resources in genomics, genetics and bioinformatics to take advantage of the post-genomic era in sunflower.

This effort could be organized in four work packages:

1. Produce genomic sequences of key sunflower genotypes
2. Develop and characterize genetic resources
3. Develop a novel genotyping tool that can assay both SNP and structural variants and genotype MAGIC population
4. Produce an atlas of stress responsive genes for genome annotation and reverse genetics

Work package 1. Produce genomic sequences of key sunflower genotypes

Task 1.1 Sequence sunflower cultivated lines using PacBio technology

Task leader: Nicolas Langlade, INRA

Partners involved: INRA, UBC, UGA

Aim: Produce the genomic sequences of sunflower lines chosen for their maximal diversity in the cultivated material. We aim to sequence 16 inbred lines representing a large genetic diversity in the cultivated pool originating both from USDA and INRA (Table 1).

Strategy: We will use PacBio sequencing based on our successful experience to sequence the XRQ and PSC8 genomes. Our strategy will be to produce very-long insert size libraries to reach an average read length of circa 14kb with a total coverage around 60X. This should represent 200 SMRTCell with a throughput of 1Gb/SMRTCell. Bioinformatics analysis will consist in *de novo* assembly and annotation for each genome. Comparison with previously sequenced genomes will be performed to aid generation of pseudomolecules.

Cost: Sequencing = 1 700 k€ (500€ / SMRTCell è 100k€ per genotype); Bioinformatics = $165 k€ ($55 k€ / year x 3 years)

Timeline: this will be spread across the first three years of the consortium. The sequencing and assembly will be performed at INRA Toulouse. Based on today's PacBio RSII sequencer outputs, we can sequence three to four sunflower genomes per platform per year. The new PacBio Sequel will allow us to sequence seven times more and therefore more easily reach our goal.

Task 1.2 Sequence wild sunflower genomes using Illumina's Long-read Sequencing Technology

Task leader: Loren Rieseberg, UBC

Partners involved: INRA, UBC, UGA

Aim: Produce the genomic sequences of wild sunflower relatives chosen for their diversity and source of resistance genes. We aim to sequence 3 *H. annuus* ecotypes carrying abiotic tolerance alleles and 6 *Helianthus* species (*H. anomalus, H. argophyllus, H. exilis, H. paradoxus, H. petiolaris* and *H. debilis*) carrying disease and orobanche resistance genes (Table 2).

Strategy: Long-read sequencing technology generates large DNA fragments for sequencing, and assembles them into synthetic long reads (circa 10kb) that can be used for assembly. Our strategy will be to produce Tru-Seq synthetic long read libraries and sequence to circa 60X depth. This approach will produce less complete assemblies than PacBio, but also at much less expense.

*Cost: Sequencing = 90 k€ ($10 k€ / genome x 9 genomes); Bioinformatics = 165 k€ ($55 k€ / year x 3 years)

Timeline: This will be spread across the first three years of the consortium. The sequencing would likely be performed at Genome Quebec in Montreal, whereas assembly would be performed at UBC.

Task 1.3 Sequence a sunflower genome using NRgene strategy

Task leader: tbd (INRA)

Partners involved: INRA, UBC, UGA

Aim: Produce the genomic sequence of a sunflower to be determined.

Strategy: We will sub-contract to NRgene. This will consist in constructing long scaffolds from short Illumina reads for the chosen sunflower genome. First NRgene will generate approx. x180 coverage of genomic DNA sequencing data. Using DeNovoMAGIC™-2 software NRgene will then construct genomic scaffolds from the data.

*Cost: Sequencing = 152k€ as quoted in February 2016
Timeline: First year

Timeline: This will be spread across the first three years of the consortium. The sequencing would likely be performed at Genome Quebec in Montreal, whereas assembly would be performed at UBC.

Work package 2. Develop and characterize genetic resources

Task 2.1. Genotyping and phenotyping of the public cultivated sunflower association mapping (SAM) population, three wild sunflower association populations that are under development, as well as public pre-bred lines.

Task leaders: Loren Rieseberg, UBC and John Burke, UGA

Partners involved: INRA, UBC, UGA

Aim: Characterize genetic polymorphisms of wild cultivated and pre-bred material

Strategy: During the first four years of the consortium, we re-sequenced the 288 cultivated genotypes making up the SAM population to 5-10x depth. This is a very large data set (>10 TB), so variant calling was a challenge. However, the SNP calls are now complete, which we will release to the consortium later this year. The SNPs were called against the bronze genome, which is not as complete or accurate as the XRQ sequence. Also, structural variants have not been called. Therefore, over the next two years, we will re-align these sequences against the XRQ reference sequence and call both SNPs and structural variants. An association mapping pipeline will be developed and provided along with phenotypic data for many agriculturally relevant traits, including drought, low nutrient, salt, and flooding tolerance. The association population has very low LD throughout much of the genome, so it is possible in many cases to map QTLs down to individual genes. Therefore, we expect this resource to be of conservable value to the sunflower community.

In addition to the SAM population resources, in the context of a new Genome Canada project, we are developing association populations for wild populations of three species: H. annuus, H. argophyllus, and H. petiolaris. Approximately 500 individuals of each species will be sequenced to 5-7x depth and their genomes will be scanned for associations with drought, low nutrient, salt, and flooding tolerance.

Lastly, in the context of an ongoing project with the Global Crop Diversity Trust on adapting crops to climate changes, we will provide phenotypic and high-resolution genotypic data for more than 400 public pre-bred lines developed by the USDA and UBC (Table 3). This will facilitate the introgression of valuable alleles from wild sunflower accessions.

*Cost: Bioinformatics = $220 k€ ($55 k€ / year x 4 years); Sequencing, genotyping, and phenotyping costs covered by NSF, Genome Canada, and Crop Trust grants

Timeline: The genotypic data for the pre-bred lines will be made available in year 1, whereas the new set of variant calls and association mapping pipeline for the SAM population with not be available until near the end of year 2. Association mapping in the wild species' association populations will be carried out in years 3 and 4. Both field- and greenhouse-derived phenotypic data will be added to the association mapping website/pipeline over the next four years.

Task 2.2. Produce a mutagenized population

Task leader: Nicolas Langlade, INRA

Partners involved: public partners + private

Aim: produce a very large mutant population in an already sequenced reference maintainer line (e.g. XRQ) for forward and reverse genetics. Produce DNA pools for screening for reverse genetics.

Strategy: Two strategies will be used: EMS and irradiation (gamma-ray) mutagenesis. The target for EMS would be 20,000 M2 families. EMS mutagenesis and production of M1 seeds can be difficult and the number of EMS families would depend on M1 seed production. Each partner would multiply 1000 families over 2 years and provide the DNA to INRA-CNRGV.

Our target for irradiation mutagenesis would be 20,000 M2 families as well. Each partner would be in charge of producing M1 and M2 generations for 2000 families. M2 seeds for both strategies will be pooled per family (>8 plants) and sent to one or several partner(s) for DNA extraction.

*Cost: Irradiation cost to be determined (quote under way), M1 and M2 multiplications by private partners. EMS: either sub-contracted or mutagenesis shared between public labs. M2 production would be shared between private partners.

20k€ for DNA extraction and 10k€ for pooling

Timeline: Irradiation strategy - mutagenesis in 2016, M1 in 2016 (if early agreement), otherwise in 2017. M2 in 2017 or 2018. EMS strategy: mutagenesis and M1 in 2017 and M2 in 2018.

Note regarding legal issues: IP for the genetic material will be shared between all WP partners. All WP partners will have access to it. Use of this genetic material in collaboration with public labs may be permitted unless a non-member company is involved.

Work package 3. Develop a novel genotyping tool that can assay both SNP and structural variants and genotype MAGIC population

Task 3.1. Develop a 500k AXIOM array for assaying both SNP and structural variants.

Task leader: Stéphane Muños, INRA

Partners involved: INRA, UBC, UGA

Aim: Produce a high-throughput genotyping tool such as Affymetrix AXIOM chip to genotype SNP and structural variants across the sunflower genome.

Strategy: identify SNPs based on resequencing data already produced by the consortium including the SAM population and genotyping by sequencing produced in WP2. The goal will be to cover every gene with at least 3 SNPs. Based on PacBio *de novo* sequencing from WP1, structural polymorphisms will be identified and markers developed to type them. Our goal is develop approximately 500k markers based on the AXIOM technology.

*Cost: 40 k€ to cover the bioinformatics design. We expect a genotyping price of circa 200€ /per sample if we have enough pre-orders. Partners involved in chip development and pre-orders will have access to the chip. Anonymous results of genotyping will be shared between partners to evaluate the tool design and marker quality.

Timeline: Chip development would take place in 2018 after most of the *de novo* sequencing was completed. We anticipate that it will take circa three months to finalize markers.

Task 3.2. Genotype MAGIC populations

Task leader: Stéphane Muños, INRA and Loren Rieseberg, UBC

Partners involved: INRA, UBC, UGA

Aim: Genotype 1200 RILs from MAGIC populations currently under development

Strategy: To facilitate the efficient genetic analysis of complex trait variation in *Helianthus* and to produce materials containing exotic alleles that can be readily deployed in breeding programs, we have partnered with industry to develop modified Multiparent Advanced Generation Inter-Cross (MAGIC) mapping populations that include both cultivated and wild sunflower donors. Such populations combine the high power to detect QTLs offered by biparental populations with the ability to assay a broader spectrum of diversity and improved resolution afforded by association mapping. Because sunflower is a hybrid crop, we are developing separate male and female populations. For each population, our crossing design involves four elite breeding lines and four wild donors, which will be intermated and selfed to develop 600 RILs. Our plan is to use the Axiom array to assay the 1200 RILS for SNP and structural variants.

*Cost: 240 k€. Of this, approximately 120€ is available for genotyping from recently funded NSF and Genome Canada projects.

Timeline: We will genotype the female MAGIC population in year 3 and the male population in year

WP 4. Produce an atlas of stress responsive genes for genome annotation and reverse genetics

Task leaders: Nicolas Langlade, INRA and John Burke, UGA

Partners involved: INRA, UBC, UGA

Aim: aid genome annotation through the identification of expressed genes and the annotation of genes as involved in tolerance to abiotic stresses and pathogens. This will be key to identifying the most likely candidate genes underlying QTLs and for targeting them for reverse genetics validation using the mutant population and/or through transformation and RNAi.

Strategy: Expression levels will be assessed by RNA-Seq. For each stress, 3-5 genotypes will be used, including reference one (e.g. XRQ), sensitive and tolerant/resistant lines. Stresses include:

- Drought stress at seedling stage: in roots and leaves at three timepoints (UGA)
- Drought stress at vegetative stress: in roots and leaves at one stage (managed on the Heliaphen platform at INRA)
- Drought stress at seed filling stage in roots, leaves and seeds at 3 stages (managed on the Heliaphen platform at INRA)
- Salt stress at seedling stage: in roots and leaves at three timepoints (UGA)
- Flooding stress at seedling stage: in roots and leaves at three timepoints (UBC)
- Low nutrient stress at seedling stage: in roots and leaves at three timepoints (UGA)
- Cold stress at seedling stage: in roots and leaves (INRA)
- Orobanche: roots infected by orobanche or not (INRA)
- Phoma: stems at the level and aound the infection by *Phoma macdonaldii* (INRA)
- Downy mildew: roots, hypocotyles and cotyledons after primary infection of *Plasmopara halstedii* (INRA)

*Cost: ~200k€ (corresponds to ~200€ per sample) for data production. Bioinformatics = $220 k€ ($55 k€ / year x 4 years); Note that ~100k€ for data production is available from NSF and Genome Canada projects.

Timeline: along the 4 years

*The costs of the project will be partly covered by consortium fees. The remaining costs will be borne by internal funds or by external grants as detailed for individual tasks.

Appendices

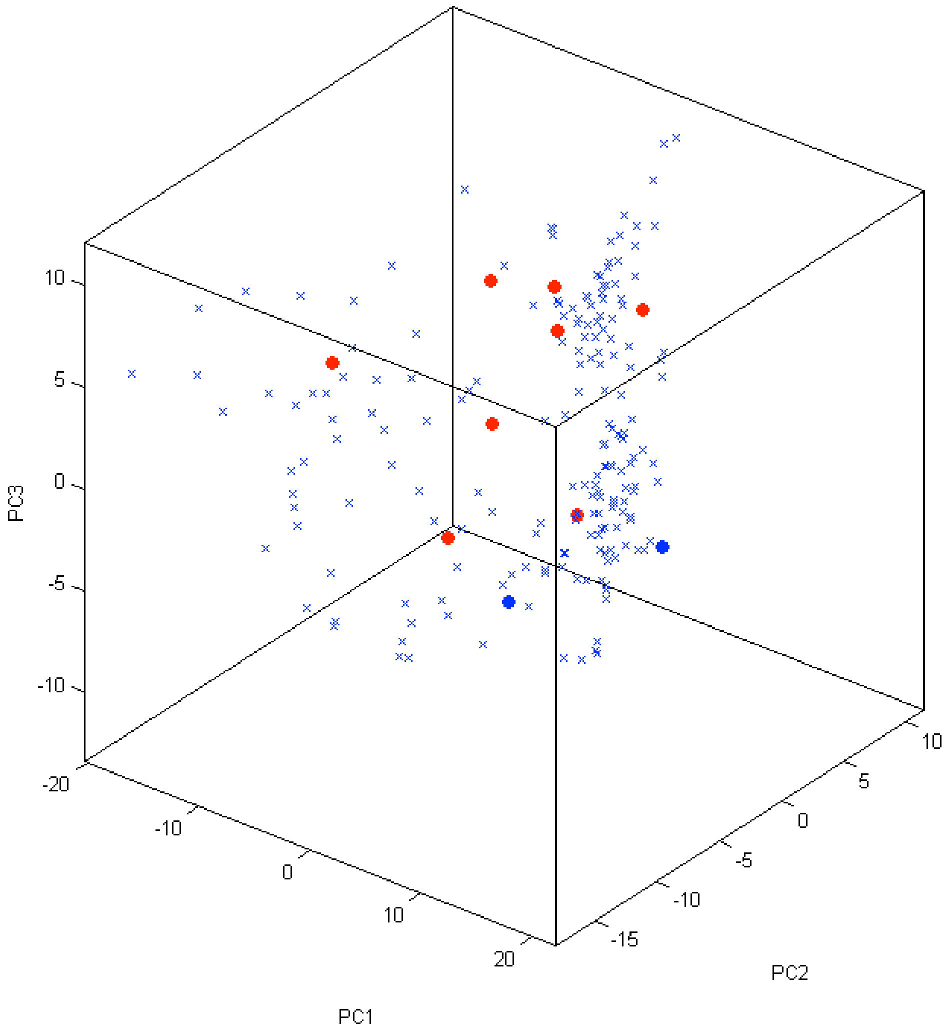

Figure 1. Principal component analysis of the genetic diversity of the public lines in the INRA core-collection using 8844 SNPs. Already sequenced lines are represented by blue circles and lines to be sequenced by PacBio are represented in red.

Table 1. Lines already sequenced (grey) and to be sequenced

Line	Other name	Source	Pedigree	Trait
SF009	ADV	INRA	ADVENT.CE.A.7.3.3.2.1.2.1.1. R100?	
SF092	IR	INRA	IMPIRA SC4°.194.8.6.C11.1.1.1.3.1	
SF109	2603	INRA	M.H.2.1..	
SF193	*XRQ*	*INRA*	*QPD(HA89 PROGR)B.1.2.7.C3.2.1.2.1*	
SF279	OQP8/OPB8	INRA	OQPCR OQP(PAA1 RHA345)A.1.9.18.1.2	
SF281	RHA438	USDA	OQPCR OQPRHA 438	

			(RHA340xRHA344 OLEIQUE)	
SF317	83HR4	INRA	PCR P(V6540 G509 RHA274) +RM+OL MTP	
SF326	*PSC8*	*INRA*	*PCR PISCMR5.313.6.7.7.5+RM*	
SF342	PW3RM	INRA	QCR PW3RM (Pl7 ex YEQ) S304	
SF230	none	INRA	USSCL 23.4.1.2(1+2)(RF-015.21.2)	Downy Mildew
RHA 358		USDA	RHA 274*3/DDR	Short stature
RHA 408		USDA	Romania R-line Sclerotinia pop-1	Sclerotinia stalk rot resistant
RHA 426		USDA	RHA 409//RHA 376*2/IMI res. H. annuus	IMISUN source, Sclerotinia stalk rot resistant
RHA 455		USDA	RHA 440/HO IS R-line	HO, Sclerotinia resistant
HA 467		USDA	HA 411/ROMPH//HA425/87CAEB/3/HA 434/HA 412	Sclerotinia and Phomopsis resistant, HO, IMISUN, moderate stature
HA 821		USDA	HA 300 selection	
HA 404		USDA	HA 821/RASSVET	Early maturity
HA89		USDA		
LR1		INRA		*Orobanche* resistance

Table 2. List of wild ecotypes to sequence using Illumina's Long-read Sequencing Technology

Species	**Population locality**	**Traits of interest**
Helianthus annuus	Western Nebraska	Drought tolerance
Helianthus annuus	Southern Manitoba, Canada	Earliness (day neutral), cold hardiness
Helianthus annuus	Great Salt Lake, Utah	Salt tolerance
Helianthus anomalus	Little Sahara Sand Dunes, Utah	Large seeds, nitrogen use efficiency
Helianthus argophyllus	Texas Coastal Plain	Drought tolerance, disease resistance
Helianthus exilis	North Coast Range, California	Heavy metal tolerance
Helianthus paradoxus	Santa Rosa, New Mexico	Salt Tolerance
Helianthus petiolaris	Great Sand Dunes National Park, Colorado	Large Seeds, nitrogen use efficiency
Helianthus debilis		Source of *Orobanche* resistance