FINAL REPORT

ConGenOmics Exchange Grant: "Exploratory genotyping-by-sequencing of elm crosses for linkage mapping and QTL detection of resistance traits to Dutch elm disease"

Beneficiary: David Macaya Sanz

Host: Erik Dahl Kjær, University of Copenhagen

1. Purpose of the visit

I proposed to explore the paths to generate linkage maps from specific crosses of resistant elms. To that aim, I explored genotyping through Restriction site Associated DNA markers sequencing (RADseq) and SSRs. Three elm parents and three offspring of each cross (in total nine offspring) were used for this exploratory genotyping.

The main objective of this stay was to identify which restriction enzymes permit to generate an adequate amount of fragments of the targeted sizes, and, afterwards which size range would permit to generate the desired number of markers to obtain a dense enough linkage map (700-1000 loci). The inclusion of SSRs was intended to compare with traditional methodologies, and to confirm which markers segregate in all the parents. In the future, it would be interesting to include traditional markers, used in many previous studies, in any linkage mapping effort. After this preparatory step, complete progenies could be RADseq genotyped with a meaningful reduction of expenses and an assurance of results.

An additional objective of my stay was to form a strong link between Erik Kjaer lab and mine, in order to create synergies to advance in the knowledge of the genomics of tree diseases.

2. Description of the work carried out during the visit

Plant material and DNA Isolation

Adult leaves (fully extended and developed) of the three parents (clones resistant to Dutch elm disease) of the three progenies and nine offspring (three per progeny) were collected and dried in silica gel during summer 2014. DNA isolation was done following a modified Doyle protocol, using ATMAB and Beta-mercaptoethanol.

Additionally, and belonging to other project, adult leaves from four more resistant clones, eight putatively pure *Ulmus minor* and eight putatively pure *Ulmus pumila* were harvested and dried in

silica gel. The results of these samples will be overviewed just to ascertain which protocol is better to isolate DNA in view of RADseq results.

Due to during the extraction of DNA appeared a large amount of mucus in several samples, especially in seedlings (i.e. offspring of the crosses), that hindered the manipulation of DNA stocks, we decided to re-isolate DNA following a different protocol. Young leaves from four offspring were then collected, dried and ground, and their DNA was isolated using a kit (Invisorb, STRATEC, Germany) with a reduced amount plant leave powder (approx. 10 mg). Also, DNA from some important adult trees whose DNA from adult leaves possessed too much mucus, in particular the three parents of the progeny, was isolated from wood tissue (phloem + xylem) using the previous kit, with 150 mg of powdered tissue. From two more adult trees, not belonging to this project, DNA was extracted from wood.

Once DNA was isolated, concentration was measured and we checked if we had the minimum amount and concentration of DNA for constructing the library. When necessary, DNA was re-concentrated by lyophilization.

All the samples were genotyped using eight microsatellites previously used in U. minor. For the time being, this information was used mainly to discard that any pair of samples were ramets of the same genet, or ramets of the *Atinio* clone, which is triploid.

Library preparation

Twenty six samples were selected for preparing the libraries and sequencing: twelve corresponding to the three crosses, and belonging to this project, and fourteen corresponding to the rest of resistant clones (4), *Ulmus minor* (5), and *U. pumila* (5).

Stock DNA was digested using two restriction enzymes: EcoRI-HF and MseI (#R3101S and #R0525S; New England Biolabs). The reaction buffer was composed by 0.02 units/µl of MseI, 0.04 units/µl of EcoRI-HF and 1x of CutSmart buffer in a total volume of 25 µl for the offspring and 50 µl for the rest of the samples. The thermal protocol was as follows: two hours at 37 °C and twenty minutes at 65 °C.

For library preparation we used the kit NEBNext DNA Library Prep Reagent Set for Illumina (#E6000S; New England Biolabs). We followed the instructions thoroughly, only halving or reducing to a quarter the amounts of digested DNA and reagents. Libraries for the parents of the progeny and the samples of the other study were processed using halved amounts of DNA and reagents. The offspring of the progeny were processed using a quarter of the proportions. The rationale of this was to evaluate if using fractions of reagents to prepare the library will produce data of enough quality, decreasing the costs of library construction two- or four-fold.

Steps of purification were done with Agencourt AMPure XP magnetic beads (Beckman Coulter) following the instructions of the library preparation kit. Size selection was also done using the magnetic beads, under instructions of the manual, targeting fragments of 320 bp (insert size 200 bp).

During the process, two offspring samples were discarded due to during the cleaning steps with magnetic beads, the beads were adsorbed in the sample solution probably by a high amount of mucus, so we were unable to finish the cleaning. Therefore, finally a total number of libraries from 24 samples were constructed: ten belonging to this project.

Fragments were amplified and labeled using the sets of primers 1 and 2 NEBNext Multiplex Oligos for Illumina (#E7355 and #E7500; New England Biolabs). Quality and amount of DNA of the libraries was assessed by means of a Nanodrop 1000 device (Thermo Scientific). A pool of the 24 samples' libraries was done by adding equal amounts of DNA from all libraries, excepting the libraries from the offspring that were halved. Thus, we expected to obtain the halved number of sequences from Illumina platform.

At the sequencing external service, the pool was filtered again to remove small fragments (probably primer dimers) and quality was checked by means of a Bioanalyzer (Agilent Technologies).

NGS runs

For the high throughput sequencing we used a whole run of a Miseq platform (Illumina), single-ended, with 150 cylces to recover sequences 150 bp long. The expected number of sequenced was twelve million. However, since the lane was shared with samples the other project, we expected approximately four million sequences for the ten samples of this project, with an expected yield of 0.6 M sequences for the parents and 0.3 for the offspring.

Bioinformatics analyses

These analyses were done with the pipeline STACKS (Catchen et al., 2013; Catchen et al., 2011). Firstly, sequences were filtered in base of reading quality using the *process_radtags* application:

Excerpt of *process_radtags* script:

```
process_radtags -p ./ -o ./samples3/ -q -r --renz_1 mseI -i gzfastq
```

Then, since for working properly with STACKS is necessary to have all the sequences of the same length, we estimated which length was the optimal one to recover the maximum of information. For that, we used the software Trimmomatic v 0.33 (Bolger et al., 2014). With this software, first we trimmed all the reads to a determined length (*CROP* option). Then, we removed all the shorter reads (*MINLEN* option). And, finally we cut the first part of the read, which were of lower quality and contained the sequence of the restriction enzymes (*HEADCROP* option).

The calculation of the sequence length threshold is not straight-forward, because when choosing very short size, then you lose a lot of information due to the trimming, but when choosing longer ones, you lose also due to read discarding, but also increase the probability that a sequence include a reading error. For our data, the optimum length was 120 bp without considering the

possible removal of sequences with errors, and 100 bp when errors are taken into account. Thus, we chose a threshold of 100 bp, removing the first 5 bp of each read:

Excerpt of Trimmomatic script:

```
java -jar trimmomatic-0.33.jar SE -phred33 MDV5_S6_L001_R1_001.fq.gz MDV5_MseI_L100.fq.gz
MINLEN:105 CROP:105 HEADCROP:5
```

Afterwards, we run the STACK pipeline for data of crosses. Firstly, we used the application *ustacks* to create the stacks of each sample. Then, we run *cstacks* to create the catalog of loci just using the information from the parents, and lastly, we contrast the value of the loci of each sample by means of *sstacks*. The settings are displayed in the following excerpt.

Excerpt of core STACKS script:

```
ustacks -t gzfastq -f MDV5_MseI_L100.fq.gz -o /MseI/ -i 1 -H -r -d -m 3 -M 3 -N 5 --max_locus_stacks 3
cstacks -b 1 -o /MseI/Cross_1/ -s MDV5_MseI_L100 -s VAD2_MseI_L100 -n 0 -p 8
sstacks -b 1 -c batch_1 - /MseI/Cross_1/ -s MDV5_MseI_L100 -s VAD2_MseI_L100 -s PH7_MseI_L100 -s
PH11_MseI_L100 -s PH13_MseI_L100  -p 8
genotypes -b 1 -P /MseI/Cross_1/ -r 1 -c
```

3. Description of the main results obtained

DNA isolation

During the DNA isolation, the main trouble found was the presence of mucus in the solution that hindered DNA quantification and library preparation.

We observed no correlation between the stock DNA concentration and the library DNA concentration (Figure 1A). Samples from seedlings, where the isolation was from adult leaves, produced more mucus. Therefore, later it was necessary to re-isolate or lyophilize to obtain the adequate DNA concentration. These samples produced a lower number of sequences, even although the DNA amount was the adequate.

NGS numbers

In total we obtained 13.67 million reads for the whole run. Of them, 3.39 million were specific to this project (10 samples). Although the library DNA amount added to the pooled library was even for all the samples, the samples that possessed higher values of library DNA concentration yielded more reads (Figure 1B). However, we observed also a light correlation between the stock DNA concentration and the number of reads obtained by each sequence (data not shown). DNA extracted from young leaves had the adequate initial DNA concentration, and produced an expected number of reads (Table 1).

Samples from adult trees with DNA extracted from wood (phloem + xylem), worked fine in general, though some samples worked poorly, especially the ones with reduced concentration of

DNA in the stock concentration. Samples from adult trees with DNA from adult leaves worked well, although none of these samples presented very high concentration of mucus.

The quality of the library DNA was not correlated with the percentage of reads retained after quality filtering (data not shown). The concentration of this DNA was only slightly correlated with the proportion of sequences retained (data not shown).

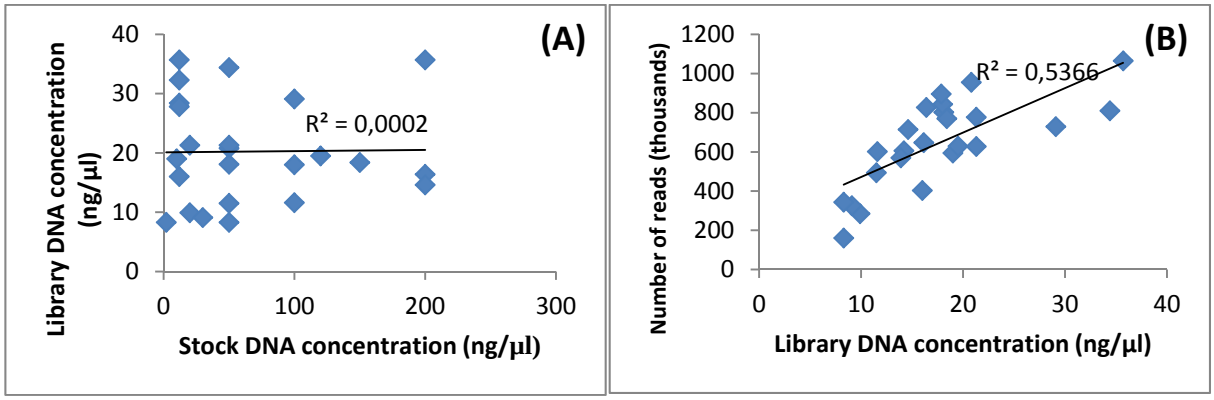

Figure 1. Correlations between the library DNA concentration and the stock DNA concentration (A) and the number of reads yielded by the NGS platform (B).

Post-processed results

The majority of the fragments began with the sequence linked to restriction enzyme MseI (85.5% of total reads after quality filtering). Reads beginning with the sequence of EcoRI were less 5%. Therefore, we focused our analysis in the first group of reads.

After trimming to length 100 bp, just 4% of reads were discarded due to short length.

Interestingly, the function linking the number of loci (or tags) detected by STACKS pipeline's *ustacks* application and the number of reads after filtering and triming did follow a potential with exponent higher than one (i.e. increasing slope) opposed to what could be expected (Figure 2A). The expectation was to find a rarefaction curve, with a decreasing slope, until reaching a tableau (i.e. horizontal asymptote). This could have been due to a reduced number of samples, although the trend is clear and robust (R^2 = 0.973), or due to the settings given to *ustacks* were the same for all the samples (see excerpt above) and they may be more strict to samples with more reads, to avoid the formation of spurious stacks. However, the number of loci created was lower than expected (25,000), probably due to the lack of a reference genome to create the stacks.

The proportion of loci created from stacks seemed to approximately converge to 0.8 (Figure 2A). This value should depend on the population genetics of the species and population examined.

After running the core STACKS pipeline with different combinations of settings, we obtained fewer loci than expected (aprox. 300 versus 700-1000; Table 2), even considering that we indicated that loci could have a lot of missing values. We acknowledge that the coverage depth seems to be insufficient, probably because we are working without a reference genome.

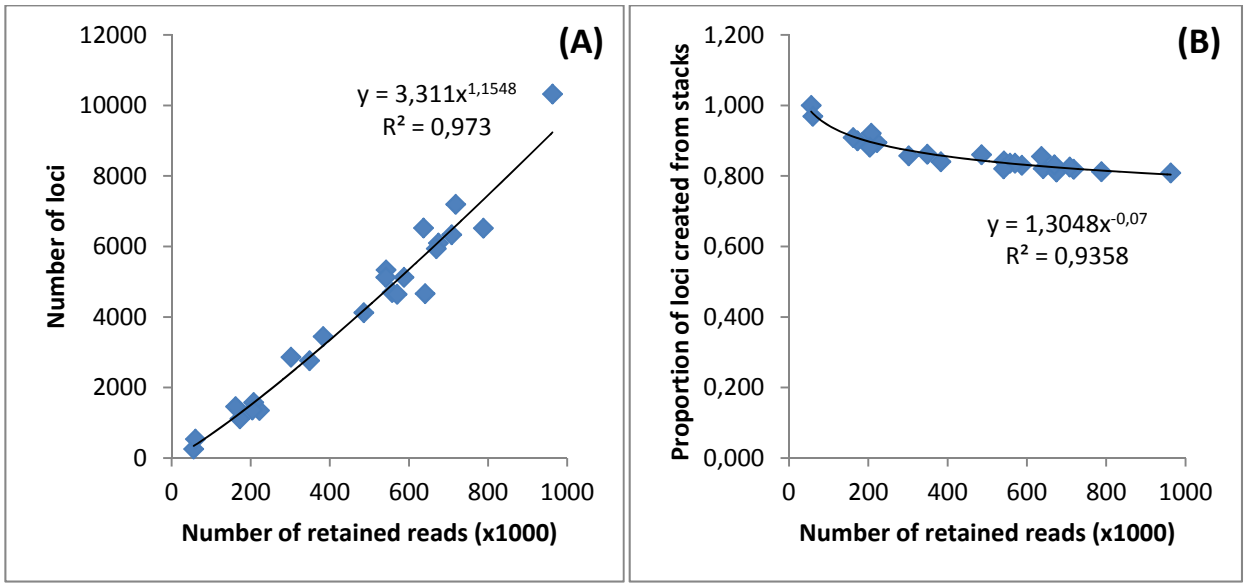

Figure 2. Correlation between the number of retained reads and the number of loci obtained by with *ustacks* (A), and the proportion of loci obtained from the stacks created (B).

Size range

Considering that the size used for the analyses has been 100 bp, but it could have been 70 bp, we find that the sequencing length could have been 120 bp, or even 100bp. However, reducing Illumina sequencing cycles to these values may be interesting only if the savings in the budget are appreciable.

Recommendations

1. To avoid overabundance of mucus, extract DNA from young leaves (not fully developed and extended), especially if working with seedlings (<10 years old). With adult trees it is less important to extract from young leaves, because they have less mucus. It is viable to extract from wood (phloem + xylem) or from adult leaves. Anyway, it is more recommendable to use young leaves.

2. Use a kit, with a small amount of powder (10 mg in case of young leaves) to avoid the formation of dense mucus. Elute the DNA to the amount of solution necessary for the next steps (measure concentration and quality, digestion with restriction enzymes). Avoid over-diluting the stock DNA.

3. Using the NEB kit for library preparation is too time-spending, especially during the steps of purification by AMPure beads. We recommend to use this protocol only in case you have few samples (<10) or you have a robot for purification in your lab. If you have many samples, it is probably better to use another method to construct the library, like the used by Lindtke et al. (2014).

4. STACKS pipeline possesses an algorithm that removes the two repetitive reads. However, with our method of size selection (AMPure beads) the distribution of fragment lengths is not uniform, but bell-shaped. Thus, STACKS could remove correct reads that are repetitive just for being in the middle of the distribution. If using STACKS, it could be interesting to

remove this algorithm or to size select fragments using another method like agarose gel band cutting.

5. Due to the poor number of reads and markers obtained and to the fall of the prices of NGS, we recommend to increase the expected number of reads per sample tenfold, to obtain approximately 3 million samples per offspring. Then, we recommend to more strict settings to create stacks avoiding the detection of spurious loci.

Approximate budget

Considering that the preparation of the library will be done using the system of New England Biolabs, used in this pilot project, we calculate that the preparation of each sample would cost 20 €. Besides, a run of a Hiseq lane costs 2500 €, and produces 350 M reads, enough for 100 samples. Each of the crosses attained in this project has 300 offspring, so the total approximate cost will be 13,500 €, without counting the in-house labor costs.

4. Future collaboration with host institution (if applicable)

We expect to continue the collaboration between the two labs in the near future. As a matter of fact, both labs are applying together for a Horizon 2020 project.

5. Projected publications / articles resulting or to result from the grant

None imminent publication is expected from this short mission. However, whenever funding is obtained to finish the genotype of the crosses, providing that these crosses display phenotypic segregation in some interesting traits, the results will be published. Besides, the samples of the other project which have been partially included in this report to give more robustness to the analyses will be analyzed and the sampling will be enlarged when funding is available. Then, another publication will be prepared. Finally, as an outcome of the interaction with other researchers in the Erik Kjaer's lab, and due to the belonging to a COST action, it is possible that a joint publication is prepared. In all these three cases, ConGenOmics funding scheme will be acknowledged.

6. References

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30**,** 2114-2120.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A., 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22**,** 3124-3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H., 2011. Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. G3-Genes Genomes Genetics 1**,** 171-182.

Lindtke, D., Gompert, Z., Lexer, C., Buerkle, C.A., 2014. Unexpected ancestry of Populus seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species. Molecular Ecology 23**,** 4316-4330.

Table 1. Summary of the sample processing, showing the samples processed (yellow: parents; green: cross 1; orange: cross 2; violet: cross 3). Note that some samples produced a reduced number of reads compared to expected, and from the too few where retained after quality filtering using process_radtags (see last two columns).

Sample	Stock DNA concentration (ng/µl)	Library preparation intial volume (µl)	Plant tissue	Plant tissue amount (mg)	Proportion in the pool	Library DNA concentration (ng/µl)	Library DNA (A260/280)	Expected number of reads (x1000)	Total number of reads (x1000)	Retained number of MseI reads (x1000)	Total/ expected	Retained/ expected	Retained/ Total
VAD2	50	45	Wood	150	2	18,1	1,71	585,37	802,96	677,52	1,37	1,16	0,84
MDV5	50	45	Wood	150	2	11,5	1,66	585,37	493,62	427,19	0,84	0,73	0,87
MRT1.5	20	45	Wood	150	2	9,9	2,03	585,37	284,66	184,71	0,49	0,32	**0,65**
PH7	12	22,5	Adult leaf	30	1	16	1,54	292,68	201,72	170,76	0,69	0,58	0,85
PH11	30	22,5	Adult leaf	30	1	9,1	1,75	292,68	162,84	62,60	0,56	0,21	**0,38**
PH13	12	22,5	Young leaf	10	1	35,7	1,94	292,68	447,81	358,64	1,53	1,23	0,80
PH4	12	22,5	Young leaf	10	1	32,3	1,86	292,68	323,79	232,27	1,11	0,79	**0,72**
PH5	2	22,5	Adult leaf	100	1	8,3	1,63	292,68	80,46	66,64	0,27	0,23	0,83
PH14	12	22,5	Young leaf	10	1	28,4	1,74	292,68	302,68	214,06	1,03	0,73	**0,71**
PH15	12	22,5	Young leaf	10	1	27,8	1,77	292,68	285,06	224,84	0,97	0,77	**0,79**

Table 2. Number of loci recovered for each of the three crosses after running the core STACKS pipeline, with different settings. First, setting the maximum distance (in nucleotides) allowed between stacks (M) to 3 or to 4 and the maximum distance allowed to align secondary reads to primary stacks to M+2. Second, setting the number of mismatches allowed between sample tags when generating the catalog to 0 or to 1. Note that the number of loci did not vary too much between settings and that it is smaller than expected (700 – 1000 bp), especially in cross 2, where a sample (PH5) yielded few reads.

	M = 3			M = 4		
Value of n	Cross 1	Cross 2	Cross 3	Cross 1	Cross 2	Cross 3
0	464	174	284	490	187	293
1	519	209	314	544	222	319