Notes on the ddRAD-seq experiment on four populations of *Coregonus sp.*

J. Ignacio Lucas Lledó

August 26, 2015

1 Motivation

After the success of the GBS experiment of *Culex pipiens*, the method is ready to be applied to other systems. There is interest in studying the history of sympatric divergent populations of *Coregonus* (fish) in a couple of lakes. In each lake there are two ecotypes (bentic and lentic). One question is whether the lentic ecotype diverged from the bentic one independently in both lakes, or if the two ecotypes are ancient lineages that colonized both lakes independently.

A fast look at the literature made me realize that simply genotyping the four populations in a series of markers will not be enough, because there is ongoing hybridization that homogenizes the neutral fraction of the variation. Thus, whether the ecological adaptations evolved twice or once, we expect the same pattern of genetic diversity: genomic islands of divergence between ecotypes in a background of undifferentiated variation [**?**].

In North America, *Coregonus clupeaformis* is present in several lakes with the two sympatric morphotypes, usually named *normal* and *dwarf*. Back in the 60s and 70s, they use isozymes and mitochondrial markers to show that they are genetically different. The first reference I find addressing the question of the origin of those genetic differences between dwarf and normal morphotypes is from 1990 [**?**]. The two hypotheses are allopatric divergence followed by secondary contact and parallel sympatric divergence. Both seem to be likely in different lakes. Across some lakes from Eastern Canada and northern Maine, two mitochondrial haplotypes A and B show clear isolation by distance, with A predominant in the East, B predominant in the West, and only three lakes with both haplotypes in the middle. Fish of clonal group B may be normal or dwarf (only in sympatry), and fish of clonal group A may also be normal or dwarf (again, dwarf only in sympatry). But in one

lake with dwarf and normal populations, the dwarf fish are all A, and the normal, B (low or null gene flux). Overall, this shows that there are two lineages diverged in allopatry that met seconarily in some lakes, where dawrf forms evolved from either one haplotype or the other.

In Northern Norway, nine lakes were surveyed for an ecomorphological and genetic study of two morphs of *Coregonus lavareuts*, here named the sparsely-rakered and the densely-rakered morphs, presumably corresponding to normal and dwarf designations [**?**]. The two morphs are well differentiated ecomorphologically, but not so by the genetic diversity of six micorsatellites. Again, a strong geographic component dominates the variation, indicating in this case that colonization of lakes proceeded from West to East (alleles in Eastern lakes are always subsets of the alleles in the West). Individuals cluster by river basin and eventually by lake, and not by morph. Given that the densely-rakered morph exists only in sympatry, it is supposed to have evolved from the sparsely-rakered morph several times independently. The study does not address the origin of the variation that determines the morph, which could be ancestral to the colonization of these lakes. Note that the Northeastern lakes were covered in ice until 9000 years ago.

2 Samples

The samples are described on table 1. Asja Vogt had already quantified them with Nanodrop, but I decided to use Qubit to have more reliable measures. Figure 1 shows the correspondance between the two methods. This time, Qubit was more optimistic.

3 Choice of restriction enzymes for double digest

During a meeting with Michael T. Monaghan, Camila Mazzoni, and Sibelle Torres Vilaça, I was advised to use double digest, instead of a single cutter. To make things easier, Sibelle offered her enzymes and adapters. She has access to P5 adapters with 5'-TA overhang (10 different barcodes), and P7 adapters with either 5'-AATT or 5'-CG overhangs (50 different barcodes). Only the P5 adapters need to have the base composition balanced, I think, because the adapters on P7 adapters are read independently. According to Sibelle, it is very important to test the enzyme cocktails and choose the right one to have a reasonable amount of fragments and enough coverage per individual. I would aim at no less than 50000 fragments.

Table 1: Samples received from Asja Vogt, and quantified with Qubit HS assay on April 13th.

Identifier	Concentration (ng/μl)	Volume (μl)	DNA mass (ng)
***Coregonus albula* (Stechlinsee)**			
St0001	29.4	99	2910.6
St0003	75.0	99	7425.0
St0006	19.0	99	1881.0
St0015	36.4	149	5423.6
St0016-2	38.4	99	3801.6
St0019-2	83.2	99	8236.8
***Coregonus fontanae* (Stechlinsee)**			
St0037-2	57.2	99	5662.8
St0039-2	48.6	99	4811.4
St0043-2	36.8	99	3643.2
St0044-2	30.6	99	3029.4
St0049-2	58.2	99	5761.8
St0050-2	33.2	99	3286.8
***Coregonus albula* (Breiter Luzin)**			
Bl0065-2	27.4	99	2712.6
Bl0076	20.8	149	3099.2
Bl0080	27.2	149	4052.8
Bl0083-2	11.1	99	1098.9
Bl0104	19.3	99	1910.7
Bl0108-2	84.2	99	8335.8
***Coregonus lucinensis* (Breiter Luzin)**			
Bl0091-2	59.8	99	5920.2
Bl0093-2	38.8	99	3841.2
Bl0094-2	49.6	99	4910.4
Bl0095-2	42.4	99	4197.6
Bl0098-2	21.4	99	2118.6
Bl0116-2	35.2	99	3484.8

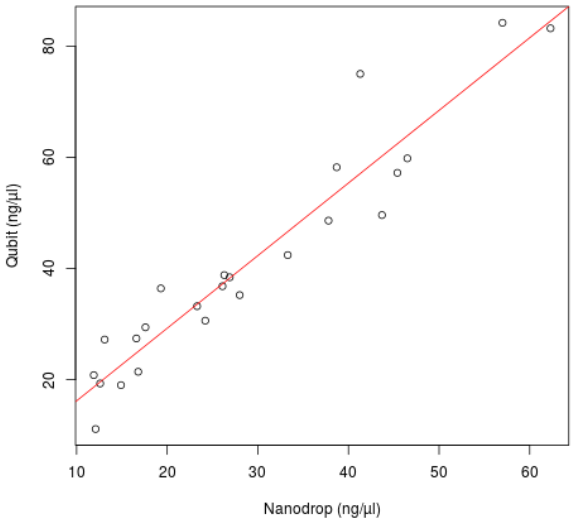

Figure 1: Relationship between the DNA concentrations determined with Qubit and those determined with Nanodrop.

In order to test the restriction enzyme combinations, I pool some DNA from the samples with highest amount of DNA (table 2).

The double restrictions will be stopped by clean-up with magnetic beads, and the product will be run through the Agilent Bioanalyzer. In order to estimate the number of copies of any fragment present, that is, the molarity of the genome, I could use an estimate of the molecular mass of one genomic copy. The Animal Genome Size Database reports an estimated haploid genome size of either 2.44 or 3.44 pg for *Coregonus clupeaformis* (depending on the reference), and 3.54 pg for *C. nasus*, the two closest relatives of *C. albula* in the database. I take the average, 3.14 pg, to represent an estimate of the genome size in *C. albula*. Using 1 pg = 978 Mb, that is 3070.92 Mb. In any case, 3.14 pg per haploid copy of the genome means that in DNA concentrations on the order of 20 ng/μl should contain around $20000/3.14 > 6000$ genome copies per μl. And total amounts of DNA shown on table 1 range between 526719.7 and 2654713.4 genome copies per sample. In order to translate concentrations or molarities of DNA fragments to number of fragments with different sequences in a size range, I use the following approach.

Consider the electropherogram produced by the Bioanalyzer to be $\Phi(x)$, in pg, keeping in mind that the amounts refer to 1 μl of sample. So, $\Phi(x)$ is the data and I model it as a product of three functions of the fragment size, x: the mass of a single double stranded DNA molecule of length x, in pg,

Table 2: Amounts of DNA from largest samples used to make a pool of DNA to test the restriction enzyme combinations. The amounts were determined proportional to the (positive) deviation from the mean of largest samples. Sibelle recommended to use at least 250 ng per reaction. I will use 300 ng.

Sample	Mass pooled (ng)	Concentration (ng/μl)	Volume pooled (μl)
St0003	209	75.0	2.8
St0015	77	36.4	2.1
St0019-2	263	83.2	3.2
St0037-2	93	57.2	1.6
St0039-2	37	48.6	0.8
St0049-2	99	58.2	1.7
Bl0108-2	269	84.2	3.2
Bl0091-2	110	59.8	1.8
Bl0094-2	43	49.6	0.9
Total	1200	66.3	18.1

Table 3: Double digest reactions.

Reagent	Volume (μl)
Frequent cutter	1.0
Rare cutter	1.0
Pooled DNA	4.5
10X CutSmart Buffer	1.8
H_2O	9.7
Total	18.0

Table 4: Definition of variables for the model of DNA fragment size distribution.

Variable	Definition
C	Mass of a haploid copy of the genome, in pg.
$\Phi(x)$	Mass of all fragments of size x, in pg.
n	Equivalent number of haploid genome copies.
$M(x)$	Mass of an individual DNA fragment of size x (pg).
$N(x)$	Number of different fragments of size x.
$R(x)$	Average redundancy of fragments of size x.

which is given by $\frac{607.4x+157.9}{6.022\times10^{23}} \times 10^{12}$; the number of *different* fragments of size x, $N(x)$; and the average redundancy among fragments of size x, $R(x)$. If the sample contained a discrete number of haploid genomes, n, the redundancy expected would be the same for every fragment: $R(x) = n$. Thus

$$\Phi(x) = \frac{nN(x)(607.4x + 157.9)}{6.022 \times 10^{11}} \tag{1}$$

where n can be estimated as $\frac{\int_0^\infty \Phi(x)\mathrm{d}x}{C}$, and $N(x)$ is the set of parameters to be estimated. I don't think it is possible to derive from that a simpler model of the mass of fragments in a size interval, with fewer parameters. That's unfortunate, because otherwise I would take full advantage of the smear analysis of 2100 Expert Software.

In any case, reactions 3 and 4 worked equally well. I choose reaction 4, which uses Time Saving enzymes, less prone to star activity, namely MseI (5'TA overhang) and MspI (5'CG overhang). These enzymes fragment the genome in pieces of on average 1045 bp,

4 Double digest

In order to minimize the chance of re-ligation of genomic fragments, which produces chimeric reads with reconstructed restriction recognition sites, Sibelle and Camila recommend to ligate the adapters in the same reaction where the genomic DNA is digested. The procedure consists on digesting first at 37°C, and then immediately add the adapters and ligase and lower the temperature. It is at this point when Sibelle adds the ligation buffer to reach 1× concentration, despite the presence of the other buffer. Sibelle insists that the samples should be kept on ice between reactions.

I have my reservations about this step, and I believe it is better to clean up the digestion reactions and measure with BioAnalyzer before ligation. Then, the adapters can be added at the right proportion with respect to ends. Since I only got about 20% of chimeras using this procedure before, I will stick to it, even though it is more work. Since I have more than 1 μg per sample, I will split the samples in more than one reaction (table 5).

After waiting for the enzymes and finding the right moment, I started digesting on June 1st. I digested the first half of samples for 1 hour at 37°C. I did not have time to clean up on the same day, so I kept the products of digestion in the fridge overnight.

5 Clean up

Double digestion reactions are stopped by clean up with magnetic beads. For clean up, all reactions from the same sample will be pooled again. Volumes will range between 381 and 596 μl. After adding 1.5 times those volumes of magnetic beads, the clean up reactions will start with between 967 and 1490 μl, which will still fit in 1500 μl tubes. That will require $1500 \times 24 \times 2 = 72000\mu$l (72 ml) of 70% ethanol. Elution volumes shown on table 6 are calculated to normalize DNA concentration at 30 ng/μl.

6 Bioanalyzer

At least two runs of the Bioanalyzer will be required to determine the molarity of fragments ends, and the expected molarity of fragments between 550 and 600 bp after ligation of adapters. With two runs, assuming both work, two samples will remail undetermined. Hopefully, some inference can be done from their DNA concentrations, based on the results for all other samples.

7 Adapters

Enzyme MseI produces 5'-TA overhangs, and enzyme MspI produces 5'-CG overhangs. Only P5 adapters, for the 5'-TA overhangs, are barcoded. The common P7 adapter, available with 5'-CG overhang gets an index after PCR. This forces me to keep samples separate until PCR. However, because size-selection is much more convenient and cheaper after pooling, I will have to run the PCR before the size-selection step. That is very unfortunate.

Table 5: Double-digestion reactions. Note that there are more than one reaction per sample, adding up to 103 reactions.

Sample	Reactions	DNA (μl)	MseI (μl)	MspI (μl)	Buffer (μl)	H_2O (μl)	Total (μl)
St0001	3	33.00	1.0	1.0	13.20	83.80	132.00
St0003	7	13.67	1.0	1.0	5.47	33.55	54.69
St0006	2	49.50	1.0	1.0	19.80	126.70	198.00
St0015	5	29.28	1.0	1.0	11.71	74.13	117.12
St0016-2	4	24.75	1.0	1.0	9.90	62.35	99.00
St0019-2	8	11.91	1.0	1.0	4.77	28.97	47.65
St0037-2	6	16.15	1.0	1.0	6.46	39.99	64.60
St0039-2	5	19.64	1.0	1.0	7.86	49.06	78.56
St0043-2	4	24.75	1.0	1.0	9.90	62.35	99.00
St0044-2	3	33.00	1.0	1.0	13.20	83.80	132.00
St0049-2	6	16.13	1.0	1.0	6.45	39.95	64.53
St0050-2	3	33.00	1.0	1.0	13.20	83.80	132.00
Bl0065-2	3	33.00	1.0	1.0	13.20	83.80	132.00
Bl0076	3	49.67	1.0	1.0	19.87	127.13	198.67
Bl0080	4	37.25	1.0	1.0	14.90	94.85	149.00
Bl0083-2	2	49.67	1.0	1.0	19.87	127.13	198.67
Bl0104	2	49.50	1.0	1.0	19.80	126.70	198.00
Bl0108-2	8	11.91	1.0	1.0	4.77	28.97	47.65
Bl 0091-2	6	16.12	1.0	1.0	6.45	39.90	64.47
Bl0093-2	4	24.75	1.0	1.0	9.90	62.35	99.00
Bl0094-2	5	19.62	1.0	1.0	7.85	49.01	78.48
Bl0095-2	4	24.75	1.0	1.0	9.90	62.35	99.00
Bl0098-2	2	49.50	1.0	1.0	19.80	126.70	198.00
Bl0116-2	3	33.00	1.0	1.0	13.20	83.80	132.00

Table 6: Clean up reactions. Elution volume is estimated to normalize concentration at 30 ng/μl.

Sample	DNA (μl)	beads (μl)	Total (μl)	Elution (μl)
St0001	396.0	594.0	990.0	97.02
St0003	382.8	574.2	957.0	239.25
St0006	396.0	594.0	990.0	62.70
St0015	585.6	878.4	1464.0	177.63
St0016-2	396.0	594.0	990.0	126.72
St0019-2	381.2	571.8	953.0	264.30
St0037-2	387.6	581.4	969.0	184.76
St0039-2	392.8	589.2	982.0	159.08
St0043-2	396.0	594.0	990.0	121.44
St0044-2	396.0	594.0	990.0	100.98
St0049-2	387.2	580.8	968.0	187.79
St0050-2	396.0	594.0	990.0	109.56
Bl0065-2	396.0	594.0	990.0	90.42
Bl0076	596.0	894.0	1490.0	103.31
Bl0080	596.0	894.0	1490.0	135.09
Bl0083-2	397.3	596.0	993.3	36.6
Bl0104	396.0	594.0	990.0	63.69
Bl0108-2	381.2	571.8	953.0	267.48
BL0091-2	386.8	580.2	967.0	192.76
BL0093-2	396.0	594.0	990.0	128.04
BL0094-2	392.4	588.6	981.0	162.19
BL0095-2	396.0	594.0	990.0	139.92
BL0098-2	396.0	594.0	990.0	70.62
BL0116-2	396.0	594.0	990.0	116.16
Total		15326.4		3356.04

This is the P5 adapter, where Xs correspond to the codeword.

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXXXX
|||||||||||||||||||||||||||||||||||||||||||
TAxxxxxxxxxxxAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

And this is the P7 adapter:

```
CGAGATCGGAAGAGCGAGAACAA
  |||||||||||||
  GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```

The products of ligation must look like this:

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTAGCAGCTAANNNNNNNNNNNNCCGAGATCGGAAGAGCGAGAACAA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGNNNNNNNNNNNNTTAGCTGCTAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

Where the codeword (variable) of the P5 adapter is marked in red, and the genomic insert, in blue. Since codewords are between 5 and 8 bases, the P5 adapter is between 40 and 43 bases. The P7 adapter (before PCR) is 34 bases on one side and 21 on the other. That is, after ligation genomic fragments are enlarged at most 79 bases. To amplify and complete the P7 flow-cell attachment sequence, a PCR is needed. One of the primers, for the P5 side of the fragment is the same as the sequencing primer:

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

The other primer includes an index (in red below), and is like this:

```
CAAGCAGAAGACGGCATACGAGATCGTTGGTTGTGACTGGAGTTCAGACGTGT
```

See below how the primers align to the fragment, and notice that the second primer can only hybridize after the first extension from primer 1.

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCT
ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTAGCAGCTAANNNNNNNNNNNNCCGAGATCGGAAGAGCGAGAACAA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGNNNNNNNNNNNNTTAGCTGCTAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
CAAGCAGAAGACGGCATACGAGATCGTTGGTTGTGACTGGAGTTCAGACGTGT
```

Thus, after the amplification PCR, the genomic fragment is up to 111 base pairs longer. Below, I represent the complete fragment and the sequencing primers for read 1, read 2 and for the index.

```
ACACTCTTTCCCTACACGACGCTCTTCCGATCT                                    GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTAGCAGCTAANNNNNNNNNNNNCCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAACCAACGATCTCGTATGCCGTCTTCTGCTTG
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CAAGCAGAAGACGGCATACGAGATCGTTGGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGNNNNNNNNNNNNTTAGCTGCTAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
```

Table 7: 100 μl of 10× annealing buffer

Component	Amount	Final concentration
Tris-HCl, pH 8, 1M	10 μl	100 mM
EDTA, 0.5 M	2 μl	10 mM
NaCl (58.44 g/mol)	2.92 mg	500 mM
H_2O	88 μl	
	100 μl	

I really want to run the size selection before the amplification PCR, so that short, undesired fragments are not present in the PCR. I can do it if I do three pools, with the up to 10 codewords in the P5 adapter, and then I use different index in the P7 adapter for each of the pools. This way the index in P7 will distinguish the pool, while the codeword in read 1 will identify the sample only within the pool. The advantage of this is that after ligation, I already can reduce the number of tubes from 24 to 3, which should make PCRs easier.

The P7 adapter needs to be annealed. Oligo P2.1 (the long one) is in solution at 200 μM, while YP2.2 is at 100 μM. I should make it into a 40 μM stock, to be at the same concentration as the P5 adapters. I estimate I will need less than 20 μl of that adapter (see table 9). I will prepare 20 μl:

1. Prepare 10× annealing buffer (see table 7). Or take some from Sibelle.
2. Combine 4 μl of 200 μM P2.1 oligo, 8 μl of 100 μM YP2.2_bio oligo, 2 μl of 10× annealing buffer, and 6 μl of nuclease-free water. Stir.
3. In a thermocycler, incubate at 97.5 °C for 2.5 minutes, and then cool at a rate not greater than 3 °C per minute, until the solution reaches a temperature of 21 °C (or lower?). Hold at 4 °C.

8 Results from the digestion and clean up

The good news is that the double digestion worked well in all samples, and the profile of fragment size distribution is very similar across all samples. The bad news is that the yield was lower than expected, around one third (table 8). Instead of using 1 μg of DNA per sample, I will use 600 μg in most samples, and only 270 ng from sample Bl083-2. Table 10 shows the actual ligation reactions.

Table 8: Amounts of DNA recovered after double digestion and clean up. These are three times the values shown in the BioAnalyzer results sheet, because the samples were diluted to one third. Values of DNA concentration for samples St0044-2 and Bl0093-2 are imputed from a linear model (see results/2015-05-05).

Sample	Original (ng)	Vol. (μl)	Conc. (ng/μl)	Final (ng)	Yield
St0001	2910.6	96.0	9.5	910.9	0.31
St0003	7177.5	238.3	11.6	2769.6	0.39
St0006	1881.0	61.7	10.9	671.6	0.36
St0015	5329.0	176.6	11.6	2045.4	0.38
St0016-2	3801.6	125.7	10.0	1255.5	0.33
St0019-2	7929.0	263.3	10.9	2880.5	0.36
St0037-2	5542.7	183.8	11.4	2088.5	0.38
St0039-2	4772.5	158.1	9.8	1555.1	0.33
St0043-2	3643.2	120.4	10.2	1222.8	0.34
St0044-2	3029.4	101.0	10.4	1046.1	0.35
St0049-2	5633.8	186.8	8.2	1530.4	0.27
St0050-2	3286.8	108.6	10.2	1108.9	0.34
Bl0065-2	2712.6	89.4	8.7	776.3	0.29
Bl0076	3099.2	102.3	8.7	891.7	0.29
Bl0080	4052.8	134.1	9.3	1249.4	0.31
Bl0083-2	1098.9	35.6	7.6	271.2	0.25
Bl0104	1910.7	62.7	10.3	643.5	0.34
Bl0108-2	8024.3	266.5	10.2	2710.7	0.34
Bl0091-2	5782.7	191.8	8.4	1601.4	0.28
Bl0093-2	3841.2	128.0	9.1	1161.6	0.30
Bl0094-2	4865.8	161.2	9.4	1516.2	0.31
Bl0095-2	4197.6	138.9	8.7	1202.6	0.29
Bl0098-2	2118.6	69.6	8.5	588.3	0.28
Bl0116-2	3484.8	115.2	12.2	1407.2	0.40

9 Ligation reactions

To design the ligation reactions, I use the estimates of DNA fragments molarity from BioAnalyzer and an estimate of the proportions of Mse-I and Msp-I compatible ends, from the *in silico* digestion of the *Salmo salar* reference genome. Among all ends, 80% are expected to be cut by MseI (TTAA), with 5'-TA overhangs and compatible with barcoded P5 adapters. The adapters will be added to the ligation reaction in 10-fold excess with respect to their corresponding fragment ends. Departing from 40 ng/μl adapter stocks, I first calculated the required dilutions to prepare a combined working stock of both adapters, each at the right concentration to produce a 10-fold excess of adapters with respect to ends in only 1 μl of the working stock. To limit the amount of adapter working stock used in the ligations is important to reduce the amount of salt added. The strategy of requiring a constant and low volume of working adapter stock per reaction is not optimal, because it requires me to prepare different stocks of the same adapter combinations for the three samples labeled with the same P5 adapter (in different pools). In view of the amounts of working stock required per sample, which are similar across samples, I decided to change the strategy and prepare only one working stock per P5 adapter (8 working stocks, instead of 24), and vary the volume of it added per reaction to accomplish the 10-fold excess of adapters in each reaction. As shown on tables 9 and 10, 14 μl of working stocks, with concentrations of P5 and P7 adapters of 23.7 and 5.7 pmol/μl respectively, contain enough adapters to create the 10-fold excess per reaction with volumes of working stock per reaction between 0.87 and 1.1 μl.

I need at least 8.3 μl of each P5 adapter, and 16 μl of the common P7 adapter.

Note that I am supplementing the ligation reactions with <50 mM NaCl, to help keep the adapters and fragments annealed, while avoiding at the same time an excess of salt that would inhibit the ligase. I will prepare fresh 1.5M NaCl solution, in case the concentration increased in the old solution, due to water evaporation. For a 1.5 M NaCl, combine 10 ml of water in a Falcon tube and 0.87 g NaCl (molecular weight 58.44 g/mol). Note that these is not necessarily nuclease-free water, which introduces a risk in the reaction. Since both the 10× T4 ligase buffer and the 40× NaCl must be added in amounts proportional to the final volume, I can pre-mix them in a 4:1 proportion and add them together to the reaction. I will make 125 μl of master mix with 100 μl 10× T4 ligase buffer and 25 μl 40× NaCl. Then the reactions will be like in tabel 11.

For the ligation reaction:

Table 9: Composition of the adapter working stocks. Both P5 and P7 adapter stocks are assumed to be at 40 ng/μl.

Samples	P5 Adapter	P5 (μl)	P7 (μl)	Annealing buffer (μl)	Total (μl)
St001, St043, Bl104	A01	8.30	2.00	3.70	14.00
St003, St044, Bl108	B01	8.30	2.00	3.70	14.00
St037, Bl080, Bl098	C01	8.30	2.00	3.70	14.00
St039, Bl083, Bl116	D01	8.30	2.00	3.70	14.00
St016, Bl065, Bl094	E01	8.30	2.00	3.70	14.00
St019, Bl076, Bl095	F01	8.30	2.00	3.70	14.00
St006, St049, Bl091	G01	8.30	2.00	3.70	14.00
St015, St050, Bl093	H01	8.30	2.00	3.70	14.00

1. Leave the DNA samples on the bench, without ice, for 5 or 10 minutes before adding the rest of reagents, to break any annealing among the overhangs of DNA fragments. Keep the working stocks of adapters on ice, to keep them annealed, and also the rest of reagents.
2. Add the adapters to the DNA samples and then put them on ice right away.
3. Add the mixture of 10× T4 ligase buffer and 40× NaCl, the water, and finally the T4 ligase.
4. Gently mix the reactions by pipetting up and down and spin them down.
5. Incubate the reactions at 16°C overnight.
6. Heat inactivate at 65°C for 10 min.
7. Allow reactions to cool down slowly (2°C per 90 s, or 0.02°C/s) to room temperature.

10 Bioanalyzer run to check ligation

After the BioAnalyzer run, the following can be said about the ligation:

Table 10: Composition of ligation reactions. Note that there are two reactions ('Num.') per sample, except for sample Bl083. 'Adapters' refers to the volume of working stock of combined P5 and P7 adapters, prepared as on table 9.

Sample	Num.	P5	Pool	DNA (μl)	Adapters (μl)	10X T4 Buffer (μl)	1.5M NaCl (μl)	T4 Ligase (μl)	Water (μl)	Total (μl)
St001	2	A01	1	31.62	0.94	4.00	1.00	2.0	0.44	40.00
St003	2	B01	1	25.81	0.98	3.30	0.82	2.0	0.08	33.00
St006	2	G01	2	27.56	1.01	3.50	0.88	2.0	0.06	35.00
St015	2	H01	2	25.91	0.98	3.40	0.85	2.0	0.86	34.00
St016	2	E01	3	30.04	0.96	3.80	0.95	2.0	0.25	38.00
St019	2	F01	3	27.42	0.99	3.50	0.88	2.0	0.22	35.00
St037	2	C01	1	26.40	0.97	3.40	0.85	2.0	0.38	34.00
St039	2	D01	1	30.50	0.95	3.90	0.97	2.0	0.67	39.00
St043	2	A01	2	29.55	0.95	3.80	0.95	2.0	0.75	38.00
St044	2	B01	2	30.00	1.01	3.80	0.95	2.0	0.24	38.00
St049	2	G01	3	36.62	0.93	4.60	1.15	2.0	0.70	46.00
St050	2	H01	3	29.37	0.95	3.70	0.93	2.0	0.06	37.00
Bl065	2	E01	1	34.56	0.93	4.30	1.07	2.0	0.14	43.00
Bl076	2	F01	1	34.42	0.96	4.30	1.07	2.0	0.24	43.00
Bl080	2	C01	2	32.20	0.87	4.10	1.02	2.0	0.80	41.00
Bl083	1	D01	2	35.64	0.91	4.50	1.12	2.0	0.83	45.00
Bl104	2	A01	3	29.23	0.97	3.70	0.93	2.0	0.17	37.00
Bl108	2	B01	3	29.49	0.98	3.80	0.95	2.0	0.78	38.00
Bl091	2	G01	1	35.92	0.94	4.50	1.12	2.0	0.51	45.00
Bl093	2	H01	1	37.50	1.10	4.70	1.18	2.0	0.52	47.00
Bl094	2	E01	2	31.89	0.97	4.00	1.00	2.0	0.14	40.00
Bl095	2	F01	2	34.66	1.00	4.40	1.10	2.0	0.84	44.00
Bl098	2	C01	3	34.81	0.95	4.40	1.10	2.0	0.74	44.00
Bl116	2	D01	3	24.55	1.07	3.20	0.80	2.0	0.38	32.00

Table 11: Composition of ligation reactions. Note that there are two reactions ('Num.') per sample, except for sample Bl083. 'Adapters' refers to the volume of working stock of combined P5 and P7 adapters, prepared as on table 9.

Sample	Num.	P5	Pool	DNA (μl)	Adapters (μl)	40× NaCl +10× buffer (μl)	T4 Ligase (μl)	Water (μl)	Total (μl)
St001	2	A01	1	31.62	0.94	5.00	2.00	0.44	40.00
St003	2	B01	1	25.81	0.98	4.12	2.00	0.08	32.99
St006	2	G01	2	27.56	1.01	4.38	2.00	0.06	35.01
St015	2	H01	2	25.91	0.98	4.25	2.00	0.86	34.00
St016	2	E01	3	30.04	0.96	4.75	2.00	0.25	38.00
St019	2	F01	3	27.42	0.99	4.38	2.00	0.22	35.01
St037	2	C01	1	26.40	0.97	4.25	2.00	0.38	34.00
St039	2	D01	1	30.50	0.95	4.87	2.00	0.67	38.99
St043	2	A01	2	29.55	0.95	4.75	2.00	0.75	38.00
St044	2	B01	2	30.00	1.01	4.75	2.00	0.24	38.00
St049	2	G01	3	36.62	0.93	5.75	2.00	0.70	46.00
St050	2	H01	3	29.37	0.95	4.63	2.00	0.06	37.01
Bl065	2	E01	1	34.56	0.93	5.37	2.00	0.14	43.00
Bl076	2	F01	1	34.42	0.96	5.37	2.00	0.24	42.99
Bl080	2	C01	2	32.20	0.87	5.12	2.00	0.80	40.99
Bl083	1	D01	2	35.64	0.91	5.62	2.00	0.83	45.00
Bl104	2	A01	3	29.23	0.97	4.63	2.00	0.17	37.00
Bl108	2	B01	3	29.49	0.98	4.75	2.00	0.78	38.00
Bl091	2	G01	1	35.92	0.94	5.62	2.00	0.51	44.99
Bl093	2	H01	1	37.50	1.10	5.88	2.00	0.52	47.00
Bl094	2	E01	2	31.89	0.97	5.00	2.00	0.14	40.00
Bl095	2	F01	2	34.66	1.00	5.50	2.00	0.84	44.00
Bl098	2	C01	3	34.81	0.95	5.50	2.00	0.74	44.00
Bl116	2	D01	3	24.55	1.07	4.00	2.00	0.38	32.00

Table 12: Total reaction volumes per sample and corresponding volumes of magnetic beads at 0.7:1 ratio. Elution volumes could be 20 μl, for a maximum final concentration of 30 ng/μl.

Sample	Ligation (μl)	Beads (μl)	Total (μl)
St0001	63.25	44.27	107.52
St0003	51.61	36.13	87.74
St0006	55.12	38.59	93.71
St0015	51.81	36.27	88.08
St0016-2	60.08	42.06	102.14
St0019-2	54.84	38.39	93.23
St0037-2	52.79	36.95	89.74
St0039-2	60.99	42.70	103.69
St0043-2	59.10	41.37	100.47
St0044-2	57.92	40.54	98.46
St0049-2	73.23	51.26	124.50
St0050-2	58.74	41.12	99.86
Bl0065-2	69.11	48.38	117.49
Bl0076	68.84	48.18	117.02
Bl0080	64.40	45.08	109.48
Bl0083-2	35.64	24.95	60.58
Bl0104	58.46	40.92	99.37
Bl0108-2	58.98	41.29	100.27
Bl0091-2	71.85	50.29	122.14
Bl0093-2	66.14	46.30	112.43
Bl0094-2	63.79	44.65	108.44
Bl0095-2	69.31	48.52	117.83
Bl0098-2	69.62	48.73	118.35
Bl0116-2	49.10	34.37	83.47

1. All samples look remarkably alike, as if the same process worked equally in all them.

2. Only fragments of 400 or more base pairs are recovered, amounting to around 28% of the original DNA.

3. The electropherograms are remarkably similar among samples, and show a simplified version of the fragment distribution, with only one peak, around 1200 bp.

4. Samples St001, St006, Bl076, and Bl094 show some additional bands, somewhat resembling some of the bands observed after digestion, but blurred in a big smear.

5. The electropherograms after ligation do not resemble the electropherograms of the same samples before ligation. Little if any coincidence of peaks is observed, even subtracting the length of the adapters from the ligated samples.

There are two possible, not exclusive, explanations of the observed change in fragment distribution. First, the low ratio of magnetic beads to DNA volume (0.7:1) had a clear effect of removing virtually all fragments below 300 bp, and probably also many of the larger fragments as well. This gradual loss of small fragments has contributed to the loss of structure (peaks) in the lower half of the distribution. On the other hand, re-ligation of genomic fragments and production of chimeric fragments is also very likely to have contributed to the loss of lower peaks. Both re-ligation of genomic fragments and depletion of short fragments are expected to produce an increase in the molarity of long fragments. But the increase caused by depletion of short fragments must be relative, rather than absulte; and the increase in molarity of large fragments should be absolute if caused by re-ligation. The question is: do I really have more large fragments after ligation? An afirmative answer is clear sign of re-ligation, while a negative one could be due to either lack of re-ligation or a combination of re-ligation and low yield.

The safest thing to do is to try again the ligation with the remaining DNA, and higher concentration of adapters, at least with the samples that have enough DNA. That could serve for comparison and eventually for back up. Plus, I have some time before the new Pippin prep comes.

11 Repeat ligation

I need to check the concentration of the adapters, and re-anneal them. Then, I will use a higher ratio of adapters to fragment ends (12-fold). I will do this

Table 13: Concentrations and remaining DNA amounts of samples analyzed in the Agilent BioAnalyzer.

Sample	Raw conc.(pg/μl)	Conc. ng/μl	DNA (ng)
St0001	2678.97	13.39	261.20
St0006	1177.39	5.89	114.80
St0039-2	2164.30	10.82	211.02
St0044-2	1728.31	8.64	168.51
St0050-2	1407.52	7.04	137.23
Bl0076	1748.27	8.74	170.46
Bl0080	1606.14	8.03	156.60
Bl0108-2	1569.44	7.85	153.02
Bl0093-2	1339.13	6.70	130.57
Bl0094-2	1914.29	9.57	186.64
Bl0116-2	1258.14	6.29	122.67

only for the samples that have still some DNA left. On table 14 I show the composition of the new working stocks of adapter combinations, and on table 15 I show the planned reactions.

12 Bioanalyzer of second ligation products

I analyzed samples St001, St003, St016, St037, St044, St050, Bl076, Bl108, Bl093, Bl094, and Bl116. Only St001, which has a very low elution volume, is clearly affected by rampant re-ligation of genomic fragments. All other samples analyzed, including Bl076 (with equally low elution volume), show a very promissing profile, which includes: a low, broad peak around 45 bp corresponging to P7 adapter dimers, a peak between 83 and 92 bp corresponding to P5 adapter dimers (peak position actually proportional to expected size: $-8.1655 + 1.1724x$, $R^2 = 0.9314$), and a complex smear that resembles the original pattern after digestion.

Since not all samples have the same pattern, I need to run the rest of samples on the Bioanalyzer to check which ones are ready for size selection, and which ones must be re-digested. Re-digestion is a suggestion from Sibelle. The idea is that if the adapters do not reconstitute the restriction site, it is a good idea to digest the samples again, in order to get rid of the chimeric fragments, which would for sure reconstitute the restriction site and be digested. This digestion does not repair the chimeric fragments, but remove them from

Table 14: Composition of the adapter working stocks for the second batch of ligations. Both P5 and P7 adapter stocks are assumed to be at 40 ng/μl.

P5	P5 (μl)	P7 (μl)	A. buf. (μl)	Total (μl)
A01	4.00	1.00	2.00	7.00
B01	4.00	1.00	2.00	7.00
C01	4.00	1.00	2.00	7.00
D01	4.00	1.00	2.00	7.00
E01	4.00	1.00	2.00	7.00
F01	4.00	1.00	2.00	7.00
G01	4.00	1.00	2.00	7.00
H01	4.00	1.00	2.00	7.00

the sequencing pool. Whether an adapter reconstitutes the restriction site or not depends on its sequence. P5 adapters were ligated to MseI-generated ends. MseI recognizes TTAA and leaves a 5'TA overhang. Thus, genomic ends will be 5'TAA...3'. P5 adapters ending in T3' would reconstitute the restriction site. The only such case is adapter H01. Thus, samples St015, St050, and St093 should NOT be re-digested with MseI. Thus, I expect a higher portion of chimeric reads, containing the TTAA pattern, in these samples.

On the other end, the short, common P7 adapter was ligated to MspI-generated ends. MspI recognizes CCGG, and it leaves 5'CG overhangs. The genomic fragments will have a 5'CGG end. The MspI recognition site would be regenerated if the P7 adapter had a C3' on the sticky side, which is not the case. Unfortunately, chimeric fragments can form by ligation of either type of ends, and MseI ends are far more frequent. I should use both enzymes when possible, and only MspI in the case of St015, St050, and St093.

The products of the second digestion that look correct will not be digested, because they indeed look very good, and even if they have some chimeras, I'm more afraid of losing the precious amount of DNA than of having a small percentage of chimeras. So, next steps:

1. Run bioanalyzer on the products of the second ligation of samples St015, St019, St039, St043, St049, Bl065, Bl080, Bl091, and Bl095, in order to have full information on what samples are ready to be processed. Because there will be two additional spots in the chip, I can also run the products of the first ligation of two samples that were not analyzed before, such as St016 and Bl095, for example.

2. Combine the products of the second ligation that need to be digested

Table 15: Composition of the second batch of ligation reactions. 'Adapters' refers to the volume of working stock of combined P5 and P7 adapters, prepared as on table 14. The mixture of 10× T4 ligase buffer and 40× NaCl can be prepared with 120 μl 10× T4 ligase buffer and 30 μl of 1.5M NaCl. Note that samples St006, Bl083, Bl104, and Bl098 are not processed here.

Sample	P5	Pool	DNA (μl)	Adapters (μl)	40× NaCl +10× buffer	Ligase (μl)	Water (μl)	Total (μl)
St001	A01	1	32.80	1.20	5.00	1.00	0.00	40.00
St003	B01	1	43.00	2.03	6.62	1.00	0.35	53.00
St015	H01	2	43.20	2.03	6.62	1.00	0.15	53.00
St016	E01	3	50.10	1.98	7.62	1.00	0.30	61.00
St019	F01	3	45.70	2.03	7.00	1.00	0.27	56.00
St037	C01	1	44.00	2.00	6.75	1.00	0.25	54.00
St039	D01	1	50.80	1.96	7.75	1.00	0.49	62.00
St043	A01	2	49.20	1.95	7.50	1.00	0.35	60.00
St044	B01	2	43.10	1.78	6.62	1.00	0.49	53.00
St049	G01	3	61.00	1.92	9.25	1.00	0.83	74.00
St050	H01	3	48.90	1.95	7.50	1.00	0.65	60.00
Bl065	E01	1	20.30	0.67	3.25	1.00	0.78	26.00
Bl076	F01	1	33.50	1.16	5.12	1.00	0.22	41.00
Bl080	C01	2	53.70	1.80	8.12	1.00	0.38	65.00
Bl108	B01	3	49.20	2.03	7.50	1.00	0.27	60.00
Bl091	G01	1	59.90	1.94	9.00	1.00	0.16	72.00
Bl093	H01	1	55.10	2.01	8.38	1.00	0.52	67.00
Bl094	E01	2	53.20	2.00	8.12	1.00	0.68	65.00
Bl095	F01	2	57.80	2.07	8.75	1.00	0.38	70.00
Bl116	D01	3	40.90	2.20	6.37	1.00	0.52	51.00

Table 16: Clean up of the products of the second batch of adapter ligations. All volumes are in μl. Elution buffer is just water, for compatibility with the Blue Pippin.

Sample	Reaction	Beads	Total	Elution
St001	40.0	32.0	72.0	13.0
St003	53.0	42.4	95.4	20.0
St015	53.0	42.4	95.4	20.0
St016	61.0	48.8	109.8	20.0
St019	56.0	44.8	100.8	20.0
St037	54.0	43.2	97.2	20.0
St039	62.0	49.6	111.6	20.0
St043	60.0	48.0	108.0	20.0
St044	53.0	42.4	95.4	18.0
St049	74.0	59.2	133.2	20.0
St050	60.0	48.0	108.0	20.0
Bl065	26.0	20.8	46.8	10.0
Bl076	41.0	32.8	73.8	12.0
Bl080	65.0	52.0	117.0	20.0
Bl108	60.0	48.0	108.0	20.0
Bl091	72.0	57.6	129.6	20.0
Bl093	67.0	53.6	120.6	20.0
Bl094	65.0	52.0	117.0	20.0
Bl095	70.0	56.0	126.0	20.0
Bl116	51.0	40.8	91.8	20.0
Total	1143.0	914.4	2057.4	373.0

with their corresponding first-ligation products (assuming all first-ligation products need to be digested).

3. Run the required digestions.
4. Stop the reaction by clean-up with magnetic beads. Use water for elution.
5. Combine the products of the second ligation that did not need to be digested with their corresponding digested counterparts.
6. Do not run another Bioanalyzer, but proceed directly to size selection with the Blue Pippin.
7. Do the Amplification PCR, quantify with Qubit, and handle the samples to the sequencing experts. Easy, eh?

13 Selective digestion of chimeric fragments

Among the products of the second batch of ligations, only samples St001 and St043 (that is, the only with adapter A01) need to be digested. They will be combined with their corresponding products of ligation 1. For the rest of samples, only the products of the first ligation need to be digested. The digestion reactions and elution volumes are on table 17. Elution buffer is water, for compatibility with the Blue Pippin.

14 Size selection

Elution volumes were kept low to ensure not much more than 30 μl per sample. I noticed a loss of volume of $\sim$2 μl during elution, after the clean up with magnetic beads. Before size selection, the volumes were corrected to at least 30 μl using water. I set up a tight range programme in the Blue Pippin machine, and selected 625 bp fragments. To date (August 6th 2015), I processed 20 of the samples, one per lane, in four cassettes, until running out of cassettes. Samples were distributed among cassettes with at least one sample of each type (lake $\times$ species) in each cassette. The last samples to be processed are St019, St050, Bl108, and Bl116. If being in water for longer is causing any problem, I expect these samples to yield lower amounts of DNA.

Table 17: Selective digestion of chimeric fragments. The DNA amounts correspond to the product of the first ligation, and in the case of samples St001 and St043 also to the product of the second ligation. Note that samples St015, St050, and St093, which have the H01 P5 adapter, are not digested with MseI, to prevent the removal of the adapters.

Sample	DNA (μl)	MspI (μl)	MseI (μl)	10X buffer (μl)	H_2O (μl)	Total (μl)	Beads (μl)	Clean up (μl)	Elution (μl)
St0001	30.5	1.00	1.00	12.20	77.30	122.00	109.80	231.80	30.00
St0003	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
St0006	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	30.00
St0015	20.0	1.00	0.00	8.00	51.00	80.00	72.00	152.00	15.00
St0016-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
St0019-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
St0037-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
St0039-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
St0043-2	38.0	1.00	1.00	15.20	96.80	152.00	136.80	288.80	30.00
St0044-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
St0049-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
St0050-2	19.5	1.00	0.00	7.80	49.70	78.00	70.20	148.20	15.00
Bl0065-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	20.00
Bl0076	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	20.00
Bl0080	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
Bl0083-2	15.0	1.00	1.00	6.00	37.00	60.00	54.00	114.00	30.00
Bl0104	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	30.00
Bl0108-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
Bl0091-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
Bl0093-2	19.5	1.00	0.00	7.80	49.70	78.00	70.20	148.20	15.00
Bl0094-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00
Bl0095-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	15.00
Bl0098-2	20.0	1.00	1.00	8.00	50.00	80.00	72.00	152.00	30.00
Bl0116-2	19.5	1.00	1.00	7.80	48.70	78.00	70.20	148.20	15.00

15 Amplification and indexing PCR

I have around 40μl (sometimes slightly less) of size-selected DNA per sample. The Phusion Master Mix is 2×, and the recommended reaction volume is up to 50 μl. I will split each sample in two, and run two PCRs per sample, with about 20 μl of DNA and 25 μl of Phusion Master Mix per reaction. That's 48 reactions, and they require 1200 μl of Phusion Master Mix.

I read in Sibelle's notebook that she uses 1 μl of 'index' (indexed primer), and 1 μl of 'IS4' in all 'Indexing PCR's. She's on vacations, but I suppose IS4 is the P5 primer. I think she makes a master mix with all common reagents, and adds the index and the DNA individually to each tube. I could use 24 (or more) different indexes, since there are plenty available. I do not have much information about the sequences of the primers. I use the same PCR programme that Sibelle has saved in her folder, but with only 15 cycles (she does 30).

Figure 2 shows that both size-selection and amplification were successful. The success in amplification implies that ligation of adapters was also successful, because primers anneal to the adapters. After 15 cycles of PCR, the shape of the fragment size distribution changes, which I interprete as a signal of differential amplification of fragments, or PCR bias. Next PCRs will run for only 10 cycles, just as in the *C. pipiens* f. pipiens samples. On table 18 I describe the amplification-indexing PCR reactions.

16 PyRAD

In parallel with the ddRAD-seq experiment with the samples from Stechlinsee and Breitzer Luzin, I am analyzing the RADseq data from [**?**], and trying to understand PyRAD. PyRAD is open source, and the code, in Python, is quite readable. It extensively uses vsearch. One frustrating feature of PyRAD is that the identity of the reads that make up the clusters and loci is kind of lost along the process. It seems to be there, in the undocumented intermediate files, but it is not easy to retrieve. For example, it would be interesting to compare the coverage across samples. However, by the time loci are identified across samples, and named, only a consensus sequence from each sample seems to be used, without reference to the number of reads used to create it. Thus, it is not possible to check the correlation of sequencing depths across samples.

I take notes here of what I understand from PyRAD's code.

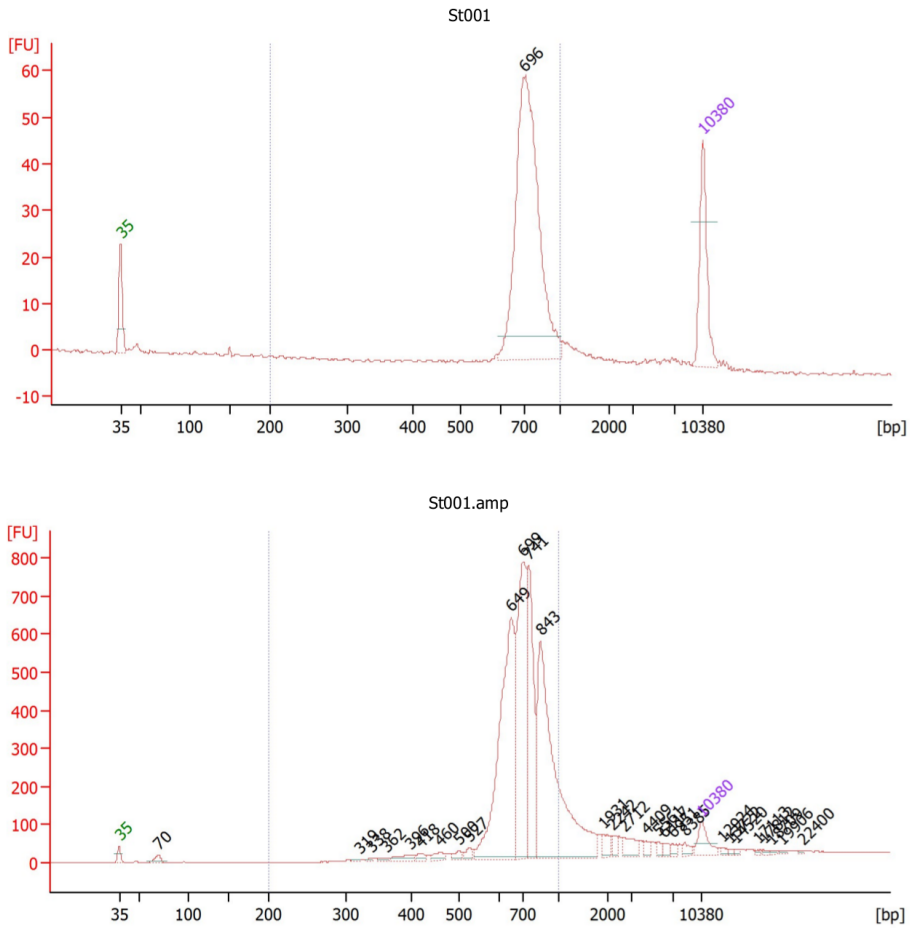

Figure 2: Amplification and indexing of 10 μl of St001 size-selected DNA with 15 cycles of PCR, using 0.5 μl of each primer.

Table 18: Amplification and indexing PCRs. Note that there are more than one reaction per sample ('Num.').

Sample	Index	Sequence	DNA (μl)	IS4 (μl)	Indexed primer (μl)	2× Phusion M/M(μl)	Cycles	Num.
St0001	index 8nt 1	AACCAACG	10.00	0.50	0.50	11.00	15	1
St0001	index 8nt 1	AACCAACG	15.00	1.50	1.50	18.00	10	2
St0003	index 8nt 129	AGAACGAC	20.00	2.00	2.00	24.00	10	2
St0006	index 8nt 200	ATAGGTAT	20.00	2.00	2.00	24.00	10	2
St0015	index 8nt 208	ATCAAGCA	20.00	2.00	2.00	24.00	10	2
St0016-2	index 8nt 229	ATTATCGA	20.00	2.00	2.00	24.00	10	2
St0019-2	index 8nt 236	CAACCTCT	20.00	2.00	2.00	24.00	10	2
St0037-2	index 8nt 256	CAGACCTT	20.00	2.00	2.00	24.00	10	2
St0039-2	index 8nt 6	AACCGGTT	10.00	1.00	1.00	12.00	15	1
St0039-2	index 8nt 6	AACCGGTT	15.00	1.50	1.50	18.00	10	2
St0043-2	index 8nt 302	CCATATAG	20.00	2.00	2.00	24.00	10	2
St0044-2	index 8nt 332	CCTGCCAA	20.00	2.00	2.00	24.00	10	2
St0049-2	index 8nt 335	CCTTGAAT	20.00	2.00	2.00	24.00	10	2
St0050-2	index 8nt 434	CTTGAGTC	20.00	2.00	2.00	24.00	10	2
Bl0065-2	index 8nt 439	GAACGCTG	20.00	2.00	2.00	24.00	10	2
Bl0076	index 8nt 526	GCTTCTCC	20.00	2.00	2.00	24.00	10	2
Bl0080	index 8nt 16	AACGGCGC	10.00	1.50	1.50	13.00	15	1
Bl0080	index 8nt 16	AACGGCGC	15.00	1.50	1.50	18.00	10	2
Bl0083-2	index 8nt 530	GGAATTGG	20.00	2.00	2.00	24.00	10	2
Bl0104	index 8nt 559	GGTCGGCG	20.00	2.00	2.00	24.00	10	2
Bl0108-2	index 8nt 582	GTCTTGGC	20.00	2.00	2.00	24.00	10	2
Bl0091-2	index 8nt 585	GTTCAATA	20.00	2.00	2.00	24.00	10	2
Bl0093-2	index 8nt 625	TCAGTAGT	20.00	2.00	2.00	24.00	10	2
Bl0094-2	index 8nt 650	TCTATTCG	20.00	2.00	2.00	24.00	10	2
Bl0095-2	index 8nt 679	TGGAGTAC	20.00	2.00	2.00	24.00	10	2
Bl0098-2	index 8nt 686	TGGTCCTG	20.00	2.00	2.00	24.00	10	2
Bl0116-2	index 8nt 711	TTGGCTCC	20.00	2.00	2.00	24.00	10	2

Step 2. Filtering. Either module editraw_rads.py or module editraw_pairs.py is called to apply some filters, and most importantly, to change low quality bases for Ns. It writes the output in edits/*.edit. Function rawedit names the sequences here, using part of the name of the input file, a number and the string 'r1', separated by underscores.

Step 3. Within-sample clustering. At this point, pyRAD calls cluster7dp.py. This first takes the 'edit' files and de-replicate them, using vsearch. It creates the files edits/*.derep. From then on, the edits/*.edit files should not be needed any more, unless the baseline quality, which determines the meaning of the threshold of low quality bases, needs to be changed. The edits/*.derep files have the multiplicity of each read added to its name, and they are ordered from higher to lower multiplicity. After de-replicating, cluster7dp.py proceeds to create the clust.XX directory and the clust.XX/*.u and clust.XX/*._temp files. These files result from the application of the cluster_smallmem algorithm with user output option. The clust.XX/*.u files are tab-delimited and they enumerate the 'query' reads, their best hits, and some customized values that qualify the match (percentage of identity, coverage of the query as portion of its length, number of gaps, and strand. The clust.XX/*._temp files hold the sequences that did not match any target. However, it is unclear what a 'query' and a 'target' are. None of the targets (second column in *.u files) is among the queries (first column); but all the targets are included in the *._temp files, as if they did not match anything. Apparently, the order in which sequences are in the input file (decreasing order of multiplicity, in this case) matters: targets have much higher multiplicity than queries. The reason must be that clusters are defined by the targets, rather than the queries. All queries that match to the same target are members of the same cluster. Thus, it is understandable that sequences with highest multiplicity (presumably, without sequencing errors) are used as centroids of the clusters. Almost all sequences in clust.XX/*._temp become a cluster, sometimes of a single sequence. Clusters are written to clust.XX/*.clustS.gz, with singletons at the end.

Step 4. Error rate and heterozygosity estimates For the maximization of the log-likelihood function, it uses the scipy.optimize module.

Step 5. Create consensus sequences The clust.XX/*.clustS.gz files are like a collection of fasta blocks, separated by `\\\n\\\n`. In step 5, the consensus function of the consensdp.py module iterates through each of those files in a very interesting way. After opening the file with gzip.open, it creates the

iterator: `k = itertools.izip(*[iter(f)]*2)`, which seems to be a duplication of the same iterator. Thus, every item in k corresponds to two lines of the file: either name and sequence from the fasta blocs, or the two separator lines. For every fasta block, it determines the consensus sequence and fills up a dictionary with the name of the first sequence (without multiplicity) as key. When consensus sequences are printed to clust.XX/*.consens.gz, they loose their original order, and all the information about the number of sequences that they come from. At least, the consensus sequences get the name of the first sequence in the cluster, which allows to retrieve the sequences that correspond to each consensus. Note that the consensus does record haplotypes (linkage among polymorphic sites within the locus) in a very smart way, by use of functions findalleles and breakalleles. When more than one heterozygous site is present, a precedence order is used (G > T > C > A) to distinguish the two alleles (upper and lower). The heterozygous sites where the upper allele is not linked to the upper allele of the first heterozygous site are encoded with lower case ambiguity codes, that can later be translated into the correct distribution of alleles between haplotypes. The number of haplotypes per sample is also checked against the assumed ploidy when more than one heterozygous site is present. Keep also in mind that building the consensus involves not only the assignment of ambiguous codes to heterozygous sites, but also it substitutes some bases for Ns (when a 3rd base is present in a site in more than 20% of reads), and remove Ns from the edges, and from within short homopolymers. Thus, the consensus sequence is shorter than its members.

Step 6. Clustering of consensus sequences This step calls the module cluster_cons7_shuf.py. It first creates a file with all consensus sequences from all samples (clust.XX/cat.group_.gz). Then, it reads it and shuffles the consensus sequences. Then it sorts the sequences by decreasing length and writes them down again in fasta format in clust.XX/cat.consens_.gz. Consensus sequences from outgroups were not shuffled, and they are added to the end of that file. For clustering purposes, the consensus sequences are converted to one of the possible haplotypes, which are stored in clust.XX/cat.haplos_. This fasta file is used as input for vsearch, which performs a clustering just as within samples. The user-defined output is clust.XX/cat.u, with a list of matches among pairs of consensus sequences. This time, target sequences are not those with higher multiplicity, but the longest ones. A dictionary is made with the names of the target sequences (second column in clust.XX/cat.u) associated with all query sequences (first column) that are similar to them. The names of sequences that belong to the same cluster are taken from that

dictionary and used to retrieve the sequences from clust.XX/cat.consens, which still has the ambiguous IUPAC codes. The clusters are written to clust.XX/cat.clust_.gz. Note that in this clustering step, sequences that do not find a match are not kept: they would be single-sample loci, which have no interest.

Step 7. Alignment, filtering of paralogs, and output files The clust.XX/cat.clust_.gz is split in chunks to be processed in parallele. Each cluster should have only one sequence from any sample, if duplicated sites are to be avoided. The number of samples represented in a cluster is also checked to skip clusters with less than a minimum. The module that performs this step is alignable.py

In conclusion, the only thing that is lost is the coverage information, which could potentially be used to detect sample-specific duplications. However, that analysis would be so sensitive to PCR biases, that is not worth attempting. Having written a script (Cor002.py) that selects the loci with a minimum number of phylogenetically informative sites, the next step is to transform that into the input format that Beast requires.

17 Pooling

Optimal DNA concentration for sequencing is 4 nM (nmol/l), and the amount should be at least 20 μl (much less is actually needed). Because of running sample-specific PCRs, I ended up with a lot of DNA. I can prepare 700 μl. I exclude sample St006, because I am not sure if the ligation work there, and I don't want to include DNA fragments that cannot be sequenced, and would only reduce sequencing efficiency. See table 19 for the proportions. I am pooling 0.12 pmols per sample, $\times$ 23 samples = 2.8 pmols; in 700 μl of water: 4 nmol/l.

Table 19:

Sample	nmol/μl	Vol. (μl)
St0001	3.48E-005	3.503
St0003	4.21E-005	2.890
St0006	3.00E-006	0.000
St0015	5.69E-005	2.140
St0016-2	6.16E-005	1.976
St0019-2	1.79E-004	0.680
St0037-2	5.19E-005	2.347
St0039-2	2.60E-005	4.680
St0043-2	4.03E-005	3.022
St0044-2	3.98E-005	3.062
St0049-2	3.29E-005	3.699
St0050-2	1.35E-004	0.900
Bl0065-2	3.42E-005	3.556
Bl0076	4.21E-005	2.890
Bl0080	2.95E-005	4.128
Bl0083-2	1.24E-005	9.795
Bl0104	2.74E-005	4.446
Bl0108-2	1.05E-004	1.162
Bl0091-2	2.36E-005	5.160
Bl0093-2	3.48E-005	3.503
Bl0094-2	3.32E-005	3.669
Bl0095-2	1.89E-005	6.457
Bl0098-2	3.95E-005	3.082
Bl0116-2	4.61E-005	2.642
Water		620.612
Total		700.000