Library scheme

We produced the library using the modified standard protocol of single digest RAD. In brief, we digested DNA by SbFI enzyme, ligated P1 (i5) adapters containing inline barcodes to each of 80 multiplexed samples, pooled 10 samples with distinct P1 barcodes (8 pooled samples), sonicated the pooled samples, ligated P2 (i7) adapters with index barcodes, pooled the samples together, amplified the fragments by PCR, and purified the library. Our final library scheme is as follows:

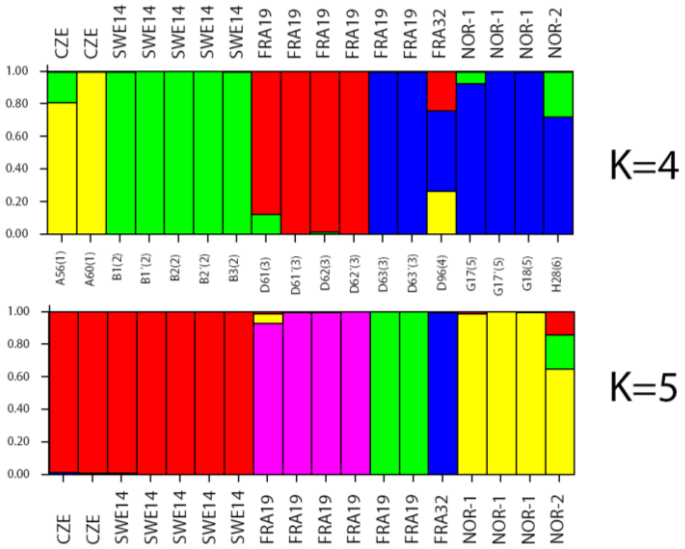
```
Final sequencing library
```

```
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXXXXTGCAGGDNA_fragmentAGATCGGAAGAGCACACGTCTGAACTCCAGTCACXXXXXXATCAGAACAA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAXXXXXXXACGTCCDNA_fragmentTTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGXXXXXXTAGAGCATACGGCAGAAGACGAAC
```

Restriction enzyme site (Sbf1-HF): CCTGCAGG
Forward PCR primer and flowcell annealing sequence: 5'-AATGATACGGCGACCACCGA-3'
Reverse PCR peimer and flowcell annealing sequence: 5'-CAAGCAGAAGACGGCATACGA-3'
Read 1 seq primer: 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
i7 index reading primer: GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'
read 2 seq primer: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'
P1 inline barcode: XXXXXXX
P2 index: XXXXXX

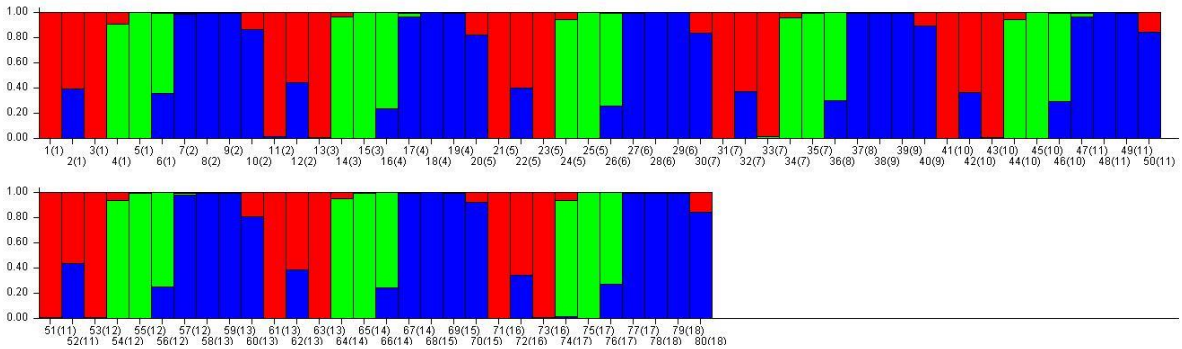
Illumina MiSeq sequencing

We successfully sequenced two libraries consisting of 9 multiplexed samples (3P1 + 3P2 barcodes) in Illumina MiSeq, using the 2x250 and 2x150 kits. The samples analysed showed a clear biogeographical pattern, i.e. they were grouped according to their origin (see the STRUCTURE plot below). Accordingly, we decided to sequence multiplexed 80 samples using Illumina HiSeq.

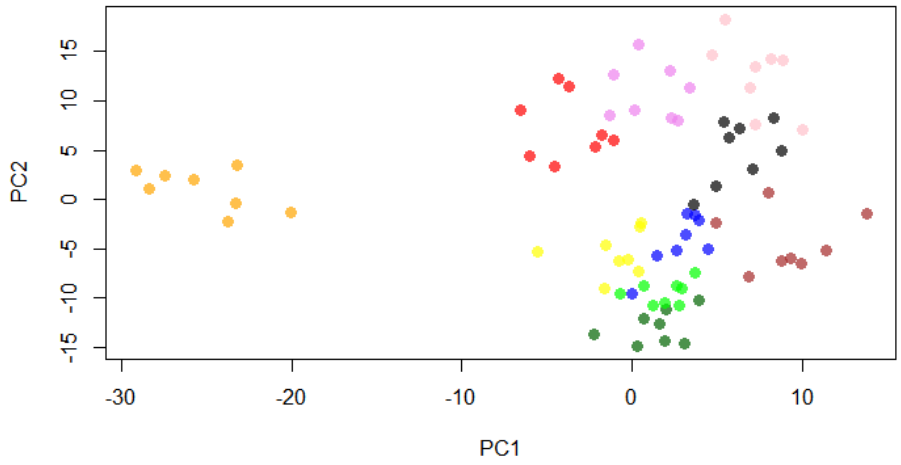


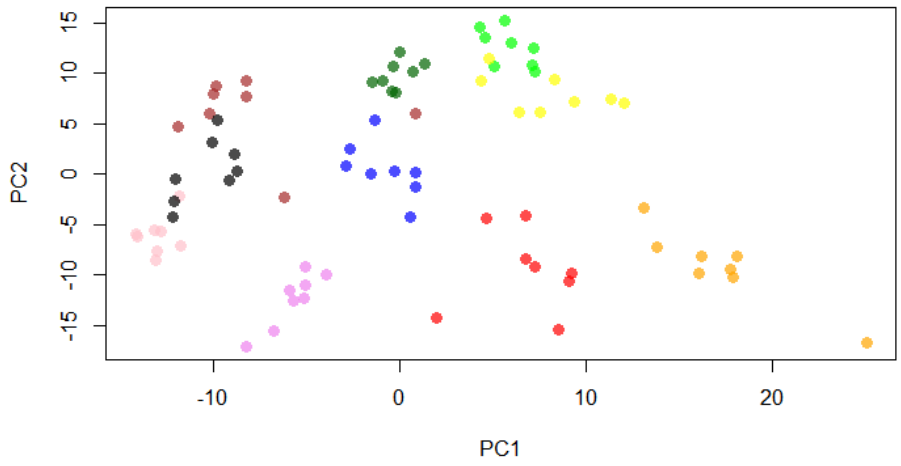
Illumina HiSeq sequencing

The first library for HiSeq sequencing consisted of 80 sequenced samples differentiated by the combinaton of 10 P1 and 8 P2 barcodes. After demultiplexing and extracting the SNPs, the samples were artificially grouped by P1 barcodes, not according to their origin. In STRUCTURE plot, there is a distinct repetitive pattern following the barcode structure. I.e., all samples with barcodes 1 and 3 were grouped to the red population, the ones with barcodes 4 and 5 to the green population, and those with barcodes 7-10 to the blue population, respectively:

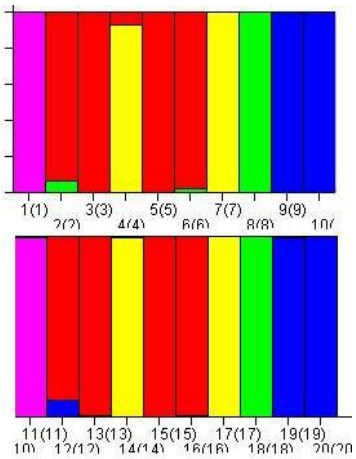


The same pattern, i.e. grouping the samples according to their barcodes, is visible in PCA plot, as well. Below are two PCA plots constructed from two sets of randomly chosen 1,300 SNPs. Samples sharing the P1 barcodes are color-coded.





Since the only explanation we had was the cross-contamination of P1 barcodes during the library construction, we constructed the second HiSeq library (Synura6), using the newly prepared barcodes, double-proofed to be free of contamination. We also changed the multiplexing strategy. Instead of multiplexing by 10 P1 and 8 P2 barcodes, we used 16 P1 and 8 P2 barcodes, to check for possible barcode swap.



After analysing the sequencing results, we found again that the samples were grouped according to the P1 barcodes. See the two STRUCTURE plots on the right, showing the similarity of those samples sharing the P1 barcode.

In addition, we detected a barcode swapping in our library (10.9% of all reads). Below is the table of read counts for all 128 barcode combinations in a 80-plex pool of our dual indexed libraries. The combinations in black text are the correct index combinations; read counts for all other combinations (in crossed cells, in grey text) are due to index swapping. The read numbers are color-coded (low numbers in red, high numbers in green). In some cases, the number of wrong combinations (swaps) exceeds the number of correct index combinations (e.g., the combinations 1-7 and 22-7).

		P2 (i7) barcodes							
		1	2	3	4	5	7	9	11
		ATCACG	CGATGT	TTAGGC	TGACCA	ACAGTG	CAGATC	GATCAG	GGCTAC
P1 (i5) barcodes	1 AACATGC	2718094	1575714	1469586	25366786	2051118	1375068	2000632	2057254
	2 AATGCCT	1369908	863988	568620	8787102	1161570	541412	1359102	839434
	3 AGAGTCG	987834	579052	249016	2832484	244216	758004	888094	378156
	4 CAATGAC	974564	630422	367008	5284846	351510	706806	811608	552278
	5 CAGACAT	1116534	461804	3088892	6926730	491458	770658	956592	779186
	8 GAGTGGA	929776	266704	1527232	4765546	282850	583342	888740	566682
	9 GCGGATA	1525706	585994	2357728	3983172	629838	918546	696884	8114122
	10 GCTTGAT	1743434	528092	2379292	3813556	570696	951498	628602	6271522
	11 GTTCAGC	961774	375932	4899572	604836	1083504	848264	556098	6045212
	15 TTCCTTC	1849050	746216	10313898	932254	1235404	1161372	861446	8282086
	16 TTCGAAG	468324	707994	5327704	504860	932950	872506	464360	4825424
	17 ACATAGG	396572	606286	3503082	474318	818124	653808	436592	5377386
	18 CGAACTG	347120	577524	1672100	396624	754172	540602	544464	5928794
	20 TCTCTCA	384090	602228	2419654	447556	721932	580346	778140	5723638
	21 CTCAGCCAAT	819316	1290234	836384	7154174	1202076	808554	1304506	9922580
	22 TGGACTTGTA	1599706	1918622	1506810	6791494	2051720	1501480	2345770	29114950

The identical results were obtained when analysing short reads only (those with overlapping pair-end reads), indicating these swaps were not formed during the PCR step in library construction. I am unable to find any explanation for the grouping of samples according to the P1 barcodes. Even if there is a barcode swap, the particular barcodes should swap randomly among the samples.