LGC Genomics GmbH
Ostendstraße 25
12459 Berlin
Germany

Tel: +49 (0)30 5304 2200
Fax: +49 (0)30 5304 2201
Email: genomics@lgcgroup.com
Web: www.lgcgroup.com/genomics

Project 12241AA-NGS777 data delivery, 28 Jul 2017:

Samples:	12
Sequencing type:	150 bp (Illumina NextSeq 500 V2)

Delivery contents:

- 'RAW': raw sequencing data after basecalling in compressed FASTQ format
- 'AdapterClipped': compressed FASTQ files containing sequencing adapter clipped reads
- 'RE_processed': compressed FASTQ files containing restriction enzyme site filtered reads
- 'QualityTrimmed': compressed FASTQ files containing quality trimmed reads
- 'Alignments': BAM formatted alignment files generated by BWA
- 'Clusters': FASTA files containing GBS clusters
- 'VariantAnalysis': Spreadsheet and VCF files containing variant calls and sample genotype data

FastQC reports (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), containing read quality metrics, are stored along with the FASTQ files.

Data analysis:

1. Read pre-processing:

- Demultiplexing of all library groups using the Illumina bcl2fastq 2.17.1.14 software (folder 'RAW', 'Group' subfolders):
    - 1 or 2 mismatches or Ns were allowed in the barcode read when the barcode distances between all libraries on the lane allowed for it
- Demultiplexing of library groups into samples according to their inline barcodes and verification of restriction site (folder 'RAW', 'Sample' subfolders)
    - no mismatches or Ns were allowed in the inline barcodes, but Ns were allowed in the restriction site
- Clipping of sequencing adapter remnants from all reads (folder 'AdapterClipped'):
    - reads with final length < 20 bases were discarded
- Restriction enzyme site filtering of read 5' ends (folder 'RE_processed'):
    - reads with 5' ends not matching the restriction enzyme site are discarded
- Quality trimming of adapter clipped Illumina reads (folder 'QualityTrimmed'):
    - removal of reads containing Ns
    - trimming of reads at 3'-end to get a minimum average Phred quality score of 20 over a window of ten bases
    - reads with final length < 20 bases were discarded
    - if one read in a pair has been discarded, the remaining mate read was written into a separate file for single reads, named *SR*

  - the R1, R2 and SR sequences were combined for alignment into one file in the 'QualityTrimmed' folder
- Subsampling (evenly across the complete FASTQ files) of quality trimmed reads to 1.5 and 3 million read pairs per sample
- Creation of FastQC reports for all FASTQ files
- Generation of read_counts.xlsx, containing all read counts for all samples at a glance

2. GBS clustering, alignment and SNP discovery:

- Clustering of combined reads with CD-HIT-EST, allowing up to 5 % difference (folder 'Clusters')
- Alignment of subsampled quality trimmed reads against all clusters using BWA version 0.7.12 (http://bio-bwa.sourceforge.net/) (folder 'Alignments'):
  - one combined alignment for all samples in coordinate-sorted BAM format
- Variant discovery and genotyping of samples with Freebayes v1.0.2-16 (https://github.com/ekg/freebayes#readme) (folder 'VariantAnalysis'):
  - the following specific parameters were used: --min-base-quality 10 --min-supporting-allele-qsum 10 --read-mismatch-limit 3 --min-coverage 5 --no-indels --min-alternate-count 4 --exclude-unobserved-genotypes --genotype-qualities --ploidy 2 --no-mnps --no-complex --mismatch-base-quality-threshold 10
- Filtering of variants using a GBS-specific rule set:
  - minimum allele count must exceed 8 reads
  - minimum allele frequency across all samples must exceed 10 % (“MinAF0.1” Excel spreadsheets)
  - genotypes must have been observed in at least 8 samples (“MinNS8” Excel spreadsheets)

If you have any questions related to your data or some steps of the data analysis, do not hesitate to contact me directly:
Email: marie.weissenborn@lgcgroup.com
Tel.: +49 (0)30 5304 2202