Sample quality report

Last update: 29.07.2021

Table of Contents

[Overview 2](#_Toc78902364)

[Arabidopsis lyrata generation 1, HM conditions 3](#_Toc78902365)

[Arabidopsis lyrata generation 4, HM conditions 5](#_Toc78902366)

[Arabidopsis lyrata generation 1, LL conditions 7](#_Toc78902367)

[Arabidopsis lyrata generation 4, LL conditions 9](#_Toc78902368)

[Arabidopsis halleri generation 1, HM conditions 11](#_Toc78902369)

[Arabidopsis halleri generation 4, HM conditions 13](#_Toc78902370)

[Arabidopsis halleri generation 1, LL conditions 15](#_Toc78902371)

[Arabidopsis halleri generation 4, LL conditions 17](#_Toc78902372)

[Arabidopsis kamchatica generation 1 synthetic, HM conditions 19](#_Toc78902373)

[Arabidopsis kamchatica generation 4 synthetic, HM conditions 22](#_Toc78902374)

[Arabidopsis kamchatica generation 1 natural (ALK), HM conditions 25](#_Toc78902375)

[Arabidopsis kamchatica generation 4 natural (ALK), HM conditions 28](#_Toc78902376)

[Arabidopsis kamchatica generation 1 natural (TKS), HM conditions 31](#_Toc78902377)

[Arabidopsis kamchatica generation 5 natural (TKS), HM conditions 34](#_Toc78902378)

[Arabidopsis kamchatica generation 1 synthetic, LL conditions 37](#_Toc78902379)

[Arabidopsis kamchatica generation 4 synthetic, LL conditions 40](#_Toc78902380)

[Arabidopsis kamchatica generation 1 natural (TKS), LL conditions 43](#_Toc78902381)

[Arabidopsis kamchatica generation 5 natural (TKS), LL conditions 46](#_Toc78902382)

[Arabidopsis kamchatica generation 1 natural (ALK), LL conditions 49](#_Toc78902383)

[Arabidopsis kamchatica generation 4 natural (ALK), LL conditions 52](#_Toc78902384)

# **Overview**

This document discusses and summarizes the main quality aspects related to the bisulfite-treated samples used to look at DNA methylation pattern changes in synthetic *Arabidopsis kamchatica*. All the aspects discussed come from the associated multiqc report (links available). The samples will be grouped in replicates and also separated according to their progenitor origin, either *halleri*-side (H-side) or *lyrata*-side (L-side). The index at the beginning of the document will help navigation for each sample.

The structure for each group of samples will follow the same format:

* Names, origins and details about the samples:   
  this general information clarifies which samples are discussed together with details about their date, platform and round of sequencing
* Trimming and general quality check:  
  trimming details and focus on potential warning issued by FastQC
* In silico conversion check:  
  results from the alignment to the *A. halleri* chloroplast genome. This is usually done to roughly assess the conversion efficiency of the samples.
* Alignment and deduplication:  
  overview of the number of aligned and deduplicated reads
* Coverage and bias:  
  considerations about the coverage and bias for each sample. Here the coverage refers to the whole genome, which might not fully capture the information needed about cytosines, but can still provide a rough idea of the starting material used to call methylation in cytosines.
* Confidence of methylation calls:  
  information digested from the output of METHImpute, assessing the confidence of all cytosines in the genome with a Hidden-Markov Model. High confidence cytosines were defined as posterior probability >= 0.9.

# ***Arabidopsis lyrata* generation 1, HM conditions**

|  |  |  |  |
| --- | --- | --- | --- |
| Species | *Arabidopsis lyrata* | | |
| Generation | Generation 1 | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | |
| Sequencing platform | NovaSeq 6000 | | |
| Date | 09.2020 | | 05.2019 |
| Sample name (short) | HM\_lyr\_G1\_1 | HM\_lyr\_G1\_2 | HM\_lyr\_G1\_3 |
| MultiQC report | HM\_syn1vpro1\_multiqc\_report.html |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. Most of the samples showed similar patterns in their per base sequence content after the second quality check, with the only exception of HM\_lyr\_G1\_3, showing slightly unexpected patterns for A/T on the R1 reads. The same samples showed some overrepresented sequences on both R1 and R2 (<1%). Sequence qualities, per base N content, duplication levels and adapter content were good after trimming. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_lyr\_G1\_1 | 6.4% | 93.6% |
| HM\_lyr\_G1\_2 | 5.5% | 94.5% |
| HM\_lyr\_G1\_3 | 2.8% | 97.2% |

The conversion error rate includes up to ~6% of all the reads in the sample. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment varied a lot between samples, highlighting differences between samples sequenced later and earlier. Mapping rates were slightly above 50% for the two samples from 2020 and at about half (~26%) for the sample from 2019, but with higher number of total reads, leading to a lower amount of uniquely mapped reads. Deduplication rates were around 5-10%, meaning that very few reads were duplicated and not too much information was lost. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_lyr\_G1\_1 | 38’693’836 | 20’259’646  (52.4%) | 19’227’517  (94.9%) |
| HM\_lyr\_G1\_2 | 27’272’244 | 14’225’747  (52.2%) | 13’604’024  (95.6%) |
| HM\_lyr\_G1\_3 | 62’028’002 | 16’624’042  (26.8%) | 15’156’295  (91.2%) |

**Coverage and bias**

Cumulative coverage varied slightly between samples for all the selected thresholds. Overall the coverage was good. M-bias plots were good for two out of three samples. For HM\_lyr\_G1\_3 some bias in the % methylation can be found in R1 reads. The bias seems to slightly underestimate methylation between 110-135bp in CpG context and between 50-135bp for CHG and CHH context. This bias doesn’t seem to largely affect downstream analyses.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_lyr\_G1\_1 | 82.5% | 77.9% | 66.4% |
| HM\_lyr\_G1\_2 | 77.0% | 63.4% | 45.4% |
| HM\_lyr\_G1\_3 | 73.4% | 63.2% | 50.9% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_lyr\_G1\_1 | High | 10'057'298 | 41'609'443 | 19.47% | 51'666'741 (94%) |
| Low | 807'451 | 2'539'070 | 24.13% | 3'346'521 (6%) |
| HM\_lyr\_G1\_2 | High | 9'830'881 | 41'232'989 | 19.25% | 51'063'870 (93%) |
| Low | 1'011'657 | 2'937'735 | 25.62% | 3'949'392 (7%) |
| HM\_lyr\_G1\_3 | High | 10'062'705 | 39'922'502 | 20.13% | 49'985'207 (91%) |
| Low | 1'322'862 | 3'705'193 | 26.31% | 5'028'055 (9%) |

For all three samples, at least 90% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis lyrata* generation 4, HM conditions**

|  |  |  |  |
| --- | --- | --- | --- |
| Species | *Arabidopsis lyrata* | | |
| Generation | Generation 4 | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | |
| Sequencing platform | NovaSeq 6000 | | |
| Date | 09.2020 | |  |
| Sample name (short) | HM\_lyr\_G4\_1 | HM\_lyr\_G4\_2 | HM\_lyr\_G4\_3 |
| MultiQC report | HM\_pro4v1\_lyr\_multiqc\_report.html | | |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. Mean quality scores, per sequence quality scores, per base N content, duplication levels, overrepresented sequences and adapter content were excellent after trimming. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_lyr\_G4\_1 | 7.3% | 92.7% |
| HM\_lyr\_G4\_2 | 8.1% | 91.9% |
| HM\_lyr\_G4\_3 | 7.7% | 92.3% |

The conversion error rate includes up to ~7-8% of all the reads in the sample. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was very similar between samples, around ~50%. Deduplication rates were around 5%, meaning that very few reads were duplicated and not too much information was lost. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_lyr\_G4\_1 | 26'258'851 | 12’893’210  (49.1%) | 12’341’584  (95.7%) |
| HM\_lyr\_G4\_2 | 30'032'166 | 14’928’542  (49.7%) | 14’146’111  (94.8%) |
| HM\_lyr\_G4\_3 | 27'601'605 | 13’735’071  (49.8%) | 13’111’848  (95.5%) |

**Coverage and bias**

Cumulative coverage was very similar between samples for all the selected thresholds. Overall the coverage was very good. M-bias plots show no striking bias in any of the samples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_lyr\_G4\_1 | 77.1% | 64.2% | 46.5% |
| HM\_lyr\_G4\_2 | 78.8% | 68.1% | 51.7% |
| HM\_lyr\_G4\_3 | 78.4% | 66.2% | 47.6% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_lyr\_G4\_1 | High | 9'582'597 | 41'175'603 | 18.88% | 50'758'200 (92%) |
| Low | 1'109'459 | 3'145'603 | 26.07% | 4'255'062 (8%) |
| HM\_lyr\_G4\_2 | High | 9'922'759 | 41'056'139 | 19.46% | 50'978'898 (93%) |
| Low | 1'028'550 | 3'005'814 | 25.50% | 4'034'364 (7%) |
| HM\_lyr\_G4\_3 | High | 9'913'788 | 41'099'561 | 19.43% | 51'013'349 (93%) |
| Low | 1'032'387 | 2'967'526 | 25.81% | 3'999'913 (7%) |

For all three samples, at least 92% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis lyrata* generation 1, LL conditions**

|  |  |  |  |
| --- | --- | --- | --- |
| Species | *Arabidopsis lyrata* | | |
| Generation | Generation 1 | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | |
| Sequencing platform | NovaSeq 6000 | | |
| Date | 09.2020 | |  |
| Sample name (short) | LL\_lyr\_G1\_1 | LL\_lyr\_G1\_2 | LL\_lyr\_G1\_3 |
| MultiQC report | LL\_syn4vpro1\_multiqc\_report.html | | |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. Mean quality scores, per sequence quality scores, per base N content, duplication levels, overrepresented sequences and adapter content were excellent after trimming. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_lyr\_G1\_1 | 6.6% | 93.4% |
| LL\_lyr\_G1\_2 | 6.5% | 93.5% |
| LL\_lyr\_G1\_3 | 5.0% | 95.0% |

The conversion error rate included 5-7% of all the reads in the sample. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was slightly over 50% for all samples, leading to an amount of uniquely mapped reads in a similar range. Deduplication rates were low, between 4-5%, meaning that very few reads were duplicated and not too much information was lost. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range, with the HM\_lyr\_G4\_3 sample having the highest number, probably because of the higher initial number of reads.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_lyr\_G1\_1 | 26'488'190 | 14’287’894  (53.9%) | 13’732’144  (96.1%) |
| LL\_lyr\_G1\_2 | 24'645'843 | 12’344’135  (50.1%) | 11’759’894  (95.3%) |
| LL\_lyr\_G1\_3 | 32'275'055 | 16’797’735  (52.0%) | 16’079’245  (95.7%) |

**Coverage and bias**

Cumulative coverage was very similar between samples for all the selected thresholds. Overall the coverage was very good. M-bias plots show no striking bias in any of the samples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_lyr\_G1\_1 | 82.1% | 77.9% | 65.5% |
| LL\_lyr\_G1\_2 | 81.3% | 75.0% | 56.7% |
| LL\_lyr\_G1\_3 | 81.1% | 73.7% | 59.8% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_lyr\_G1\_1 | High | 9'946'851 | 40'853'383 | 19.58% | 50'800'234 (92%) |
| Low | 1'198'592 | 3'014'436 | 28.45% | 4'213'028 (8%) |
| LL\_lyr\_G1\_2 | High | 9'956'989 | 40'463'028 | 19.75% | 50'420'017 (92%) |
| Low | 1'402'759 | 3'190'486 | 30.54% | 4'593'245 (8%) |
| LL\_lyr\_G1\_3 | High | 9'840'686 | 41'429'008 | 19.19% | 51'269'694 (93%) |
| Low | 912'808 | 2'830'760 | 24.38% | 3'743'568 (7%) |

# ***Arabidopsis lyrata* generation 4, LL conditions**

|  |  |  |
| --- | --- | --- |
| Species | *Arabidopsis lyrata* | |
| Generation | Generation 4 | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | |
| Sequencing platform | NovaSeq 6000 | |
| Date | 01.2020 |  |
| Sample name (short) | LL\_lyr\_G4\_1 | LL\_lyr\_G4\_2 |
| MultiQC report | LL\_lyr4vlyr1\_multiqc\_report.html | |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. Mean quality scores, per sequence quality scores, per base N content, duplication levels, overrepresented sequences and adapter content were excellent after trimming. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_lyr\_G4\_1 | 6.2% | 93.8% |
| LL\_lyr\_G4\_2 | 5.5% | 94.5% |

The conversion error rate included 5-6% of all the reads in the sample. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was slightly less than 50% for both samples, leading to an amount of uniquely mapped reads in a similar range. Deduplication rates were around 7-8%, meaning that very few reads were duplicated and not too much information was lost. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_lyr\_G4\_1 | 43'519'949 | 21’136’005  (48.6%) | 19’618’831  (92.8%) |
| LL\_lyr\_G4\_2 | 35'040'451 | 16’499’582  (47.1%) | 15’437’368  (93.6%) |

**Coverage and bias**

Cumulative coverage was very similar between samples for all the selected thresholds, with LL\_lyr\_G1\_2 showing a larger decrease in coverage for the 15X threshold. Overall the coverage was very good. M-bias plots show no striking bias in any of the samples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_lyr\_G1\_1 | 82.2% | 78.0% | 68.6% |
| LL\_lyr\_G1\_2 | 81.4% | 75.7% | 56.7% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_lyr\_G4\_1 | High | 10'489'600 | 40'944'366 | 20.39% | 51'433'966 (93%) |
| Low | 853'514 | 2'725'782 | 23.85% | 3'579'296 (7%) |
| LL\_lyr\_G4\_2 | High | 10'363'643 | 40'746'809 | 20.28% | 51'110'452 (93%) |
| Low | 1'034'810 | 2'868'000 | 26.51% | 3'902'810 (7%) |

For all three samples, at least 93% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis halleri* generation 1, HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis halleri* | | | |
| Generation | Generation 1 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 09.2020 | | | 05.2019 |
| Sample name (short) | HM\_hal\_G1\_1 | HM\_hal\_G1\_2 | | HM\_hal\_G1\_3 |
| MultiQC report | HM\_syn1vpro1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. All of the samples but one showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapter. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion. For the sample HM\_hal\_G1\_3, after trimming, quality scores were low between 110-135bp and per base sequence content showed some bias in %T and %A in the 0-30bp part.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_hal\_G1\_1 | 3.8% | 96.2% |
| HM\_hal\_G1\_2 | 5.1% | 94.9% |
| HM\_hal\_G1\_3 | 2.7% | 97.3% |

The conversion error rate was between 3-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was slightly over 50% for two out of three samples, with HM\_hal\_G1\_3 having a lower mapping rate but a higher number of total reads. Deduplication rates were lower for two samples, between 5-6%, and slightly higher in HM\_hal\_G1\_3 with ~8%. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range, with HM\_hal\_G1\_3 having the least reads, but still not too far from the other samples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_hal\_G1\_1 | 30'523'042 | 16’501’774  (54.1%) | 15’736’120  (95.4%) |
| HM\_hal\_G1\_2 | 24'216'644 | 12’137’354  (50.1%) | 11’643’182  (95.9%) |
| HM\_hal\_G1\_3 | 40'168'222 | 10’821’165  (26.9%) | 9’998’430  (92.4%) |

**Coverage and bias**

Cumulative coverage varied between samples for all the selected thresholds. Overall the coverage was good for two out of three samples, with HM\_hal\_G1\_3 showing ~1/3 less coverage on all the selected thresholds. M-bias plots were good for two out of three samples. For HM\_hal\_G1\_3 some bias in the % methylation can be found in R1 reads. The bias seems to slightly underestimate methylation in CHG and CHH context. This bias doesn’t seem to largely affect downstream analyses.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_hal\_G1\_1 | 76.2% | 61.9% | 47.4% |
| HM\_hal\_G1\_2 | 77.9% | 68.8% | 47.5% |
| HM\_hal\_G1\_3 | 60.6% | 45.3% | 30.4% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_hal\_G1\_1 | High | 11'814'839 | 43'512'381 | 21.35% | 55'327'220 (92%) |
| Low | 1'309'705 | 3'521'143 | 27.11% | 4'830'848 (8%) |
| HM\_hal\_G1\_2 | High | 12'158'109 | 42'708'415 | 22.16% | 54'866'524 (91%) |
| Low | 1'669'434 | 3'622'110 | 31.55% | 5'291'544 (9%) |
| HM\_hal\_G1\_3 | High | 11'321'706 | 41'795'462 | 21.31% | 53'117'168 (88%) |
| Low | 1'964'892 | 5'076'008 | 27.91% | 7'040'900 (12%) |

For all three samples, at least ~90% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis halleri* generation 4, HM conditions**

|  |  |  |  |
| --- | --- | --- | --- |
| Species | *Arabidopsis halleri* | | |
| Generation | Generation 4 | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | |
| Sequencing platform | NovaSeq 6000 | | |
| Date | 09.2020 | |  |
| Sample name (short) | HM\_hal\_G4\_1 | HM\_hal\_G4\_2 | HM\_hal\_G4\_3 |
| MultiQC report | HM\_pro4v1\_hal\_multiqc\_report.html | | |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. Mean quality scores, per sequence quality scores, per base N content, duplication levels, overrepresented sequences and adapter content were excellent after trimming. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_hal\_G4\_1 | 5.1% | 94.9% |
| HM\_hal\_G4\_2 | 4.7% | 95.3% |
| HM\_hal\_G4\_3 | 5.1% | 94.9% |

The conversion error rate was between 4-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was very similar between samples, between 50-52%. Deduplication rates were around 4%, meaning that very few reads were duplicated and not too much information was lost. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_hal\_G4\_1 | 32'275'713 | 16’250’034  (50.3%) | 15’596’898  (96.0%) |
| HM\_hal\_G4\_2 | 27'630'244 | 14’512’857  (52.5%) | 13’943’697  (96.1%) |
| HM\_hal\_G4\_3 | 25'440'966 | 13’375’694  (52.6%) | 12’787’398  (95.6%) |

**Coverage and bias**

Cumulative coverage was very similar between samples for all the selected thresholds. HM\_hal\_G1\_1 coverage values were highest for all thresholds, followed by 2 and then 3. Overall the coverage was very good. M-bias plots show no striking bias in any of the samples.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_hal\_G1\_1 | 77.7% | 68.6% | 52.2% |
| HM\_hal\_G1\_2 | 74.4% | 60.7% | 42.9% |
| HM\_hal\_G1\_3 | 70.0% | 54.6% | 38.4% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_hal\_G4\_1 | High | 11'465'989 | 43'983'706 | 20.68% | 55'449'695 (92%) |
| Low | 1'323'417 | 3'384'956 | 28.11% | 4'708'373 (8%) |
| HM\_hal\_G4\_2 | High | 11'315'025 | 44'029'280 | 20.44% | 55'344'305 (92%) |
| Low | 1'362'383 | 3'451'380 | 28.30% | 4'813'763 (8%) |
| HM\_hal\_G4\_3 | High | 11'265'896 | 43'659'463 | 20.51% | 54'925'359 (91%) |
| Low | 1'474'385 | 3'758'324 | 28.18% | 5'232'709 (9%) |

For all three samples, at least 91% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis halleri* generation 1, LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis halleri* | | | |
| Generation | Generation 1 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 09.2020 | | |  |
| Sample name (short) | LL\_hal\_G1\_1 | LL\_hal\_G1\_2 | | LL\_hal\_G1\_3 |
| MultiQC report | LL\_hal4vhal1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_hal\_G1\_1 | 4.7% | 95.3% |
| LL\_hal\_G1\_2 | 4.1% | 95.9% |
| LL\_hal\_G1\_3 | 4.1% | 95.9% |

The conversion error rate was between 4-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was around 55% for all samples and both total and uniquely mapped reads had values in a very close range. Deduplication rates were 5% or lower for all samples, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_hal\_G1\_1 | 29'337'779 | 16’243’449  (55.4%) | 15’632’380  (96.3%) |
| LL\_hal\_G1\_2 | 30'975'795 | 17’325’326  (55.9%) | 16’455’804  (95.0%) |
| LL\_hal\_G1\_3 | 29'894'289 | 16’256’745  (54.4%) | 15’519’348  (95.5%) |

**Coverage and bias**

Cumulative coverage was very similar between samples for all the selected thresholds. Overall the coverage was good for all samples and no worrisome pattern was found. M-bias plots were good for all three samples, showing no under or overestimation of the methylation %.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_hal\_G1\_1 | 78.4% | 70.5% | 55.2% |
| LL\_hal\_G1\_2 | 76.9% | 67.0% | 51.5% |
| LL\_hal\_G1\_3 | 75.5% | 63.7% | 47.3% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_hal\_G1\_1 | High | 12312631 | 43291298 | 22.14% | 55'603'929 (92%) |
| Low | 1288339 | 3265800 | 28.29% | 4'554'139 (8%) |
| LL\_hal\_G1\_2 | High | 12186361 | 43282307 | 21.97% | 55'468'668 (92%) |
| Low | 1328534 | 3360866 | 28.33% | 4'689'400 (8%) |
| LL\_hal\_G1\_3 | High | 9840686 | 41429008 | 19.19% | 51'269'694 (92%) |
| Low | 1310318 | 3340783 | 28.17% | 4'651'101 (8%) |

For all three samples, ~92% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis halleri* generation 4, LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis halleri* | | | |
| Generation | Generation 4 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 01.2020 | | |  |
| Sample name (short) | LL\_hal\_G4\_1 | LL\_hal\_G4\_2 | |  |
| MultiQC report | LL\_hal4vhal1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_hal\_G4\_1 | 5.3% | 94.7% |
| LL\_hal\_G4\_2 | 4.8% | 95.2% |

The conversion error rate was around 5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after alignment was around 50% for both samples. Deduplication rates around 6% for all samples, meaning that limited reads were lost in the process. Overall, the final number of reads after alignment and deduplication ended up in a relatively close range.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Total reads (pairs) | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_hal\_G4\_1 | 36'566'107 | 18’013’302  (49.3%) | 16’966’990  (94.2%) |
| LL\_hal\_G4\_2 | 31'576'035 | 15’703’201  (49.7%) | 14’770’774  (94.1%) |

**Coverage and bias**

Cumulative coverage was similar between samples for all the selected thresholds, with LL\_hal\_G4\_2 showing ~10% less in 15X coverage. Overall the coverage was good for all samples and no worrisome pattern was found. M-bias plots were good for all three samples, showing no under or overestimation of the methylation %.

|  |  |  |  |
| --- | --- | --- | --- |
|  | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_hal\_G4\_1 | 79.6% | 75.4% | 63.9% |
| LL\_hal\_G4\_2 | 77.4% | 68.5% | 52.3% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_hal\_G4\_1 | High | 12'312'631 | 43'291'298 | 22.14% | 55'603'929 (92%) |
| Low | 1'288'339 | 3'265'800 | 28.29% | 4'554'139 (8%) |
| LL\_hal\_G4\_2 | High | 12'186'361 | 43'282'307 | 21.97% | 55'468'668 (92%) |
| Low | 1'328'534 | 3'360'866 | 28.33% | 4'689'400 (8%) |

For all three samples, ~92% of cytosines’ methylation state was called with high confidence. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 synthetic, HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 01.2020 | | | 01.2019 |
| Sample name (short) | HM\_RS7K\_G1\_2 | HM\_RS7K\_G1\_3 | | HM\_RS7K\_G1\_1 |
| MultiQC report | HM\_syn1vpro1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good per sequence quality scores, per base N content, duplication levels, overrepresented sequences and no adapters. Two out of three samples had good quality scores, with HM\_RS7K\_G1\_1 showing lower scores between 110-135bp. In addition, the R1 of HM\_RS7K\_G1\_1 showed a skewed per base sequence content, particularly for %T. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_RS7K\_G1\_1 | 2.4% | 97.6% |
| HM\_RS7K\_G1\_2 | 4.8% | 95.2% |
| HM\_RS7K\_G1\_3 | 4.4% | 95.6% |

The conversion error rate was between 2-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was quite different between samples. For HM\_RS7K\_G1\_1 the total number of reads was quite high and the mapping rate was lower than the other two samples, leading to ~40% less uniquely aligned reads. HM\_RS7K\_G1\_2 and 3 were quite similar in their alignment rate, with the exception of HM\_RS7K\_G1\_2 *lyrata* side, having ~8% lower alignment rate. Deduplication rates were around 7-8%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for two out of three samples, with HM\_RS7K\_G1\_1 requiring potential attention downstream.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_RS7K\_G1\_1 | 111'829'503 | hal | 18’891’340  (16.9%) | 17’374’717  (92.0%) |
| lyr | 18’020’504  (16.1%) | 16’580’386  (92.0%) |
| HM\_RS7K\_G1\_2 | 82'499'019 | hal | 30’179’793  (36.6%) | 28’212’460  (93.5%) |
| lyr | 23’844’961  (28.9%) | 22’288’984  (93.5%) |
| HM\_RS7K\_G1\_3 | 73'276'049 | hal | 25’573’510  (34.9%) | 23’672’569  (92.6%) |
| lyr | 25’204’786  (34.4%) | 23’330’309  (92.6%) |

**Coverage and bias**

Cumulative coverage was similar for two out of three samples for all the selected thresholds, with HM\_RS7K\_G1\_1 showing ~10-15% less in all coverages for both sides. M-bias plots were good for all two out of three samples, with HM\_RS7K\_G1\_1 showing some skewed pattern for R1 reads in CHG and CHH context, both slightly overestimating % methylation in the first half of the read and underestimating in the second half. This biased estimation is difficult to quantify as it falls within the range of the other samples but instead of being constant, it constantly decreases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_RS7K\_G1\_1 | hal | 67.9% | 56.1% | 42.8% |
| lyr | 71.4% | 60.9% | 48.0% |
| HM\_RS7K\_G1\_2 | hal | 78.5% | 74.8% | 67.9% |
| lyr | 76.1% | 69.2% | 57.3% |
| HM\_RS7K\_G1\_3 | hal | 76.9% | 70.2% | 58.6% |
| lyr | 80.5% | 75.7% | 66.2% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_RS7K\_G1\_1 | High | 11'676'718 | 41'851'266 | 21.81% | 53'527'984 (89%) |
| Low | 1'752'307 | 4'877'777 | 26.43% | 6'630'084 (11%) |
| HM\_RS7K\_G1\_2 | High | 12'333'230 | 43'136'754 | 22.23% | 55'469'984 (92%) |
| Low | 1'225'714 | 3'462'370 | 26.15% | 4'688'084 (8%) |
| HM\_RS7K\_G1\_3 | High | 11'964'181 | 42'940'948 | 21.79% | 54'905'129 (91%) |
| Low | 1'395'265 | 3'857'674 | 26.56% | 5'252'939 (9%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_RS7K\_G1\_1 | High | 9'523'420 | 39'892'195 | 19.27% | 49'415'615 (90%) |
| Low | 1'267'571 | 4'330'076 | 22.64% | 5'597'647 (10%) |
| HM\_RS7K\_G1\_2 | High | 9'687'344 | 39'491'294 | 19.70% | 49'178'638 (89%) |
| Low | 1'115'257 | 4'719'367 | 19.11% | 5'834'624 (11%) |
| HM\_RS7K\_G1\_3 | High | 9'845'074 | 41'306'756 | 19.25% | 51'151'830 (93%) |
| Low | 919'542 | 2'941'890 | 23.81% | 3'861'432 (7%) |

For all three samples, ~ 90% of cytosines’ methylation state was called with high confidence. For most samples, low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 4 synthetic, HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 4 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 08.2021 | | |  |
| Sample name (short) | HM\_RS7\_G4\_1 | HM\_RS7\_G4\_2 | | HM\_RS7\_G4\_3 |
| MultiQC report | HM\_syn4vpro1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_RS7\_G4\_1 | 5.7% | 94.3% |
| HM\_RS7\_G4\_2 | 5.3% | 94.7% |
| HM\_RS7\_G4\_3 | 6.9% | 93.1% |

The conversion error rate was between 5-7%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed differences between sides of about 10%. Specifically, the lyr side showed lower mapping rates compared to the hal side. Deduplication rates were around 5-6%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for this slight imbalance between subgenomes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_RS7\_G4\_1 | 74'457'409 | hal | 31'649'347  (42.5%) | 28336892  (89.6%) |
| lyr | 22'808'665  (30.6%) | 20425020  (89.6%) |
| HM\_RS7\_G4\_2 | 75'226'540 | hal | 26'482'214  (35.2%) | 24'589'153  (92.9%) |
| lyr | 18'523'123  (24.6%) | 17'202'072  (92.9%) |
| HM\_RS7\_G4\_3 | 80'065'803 | hal | 33'614'361  (42.0%) | 30'886'272  (91.9%) |
| lyr | 25'719'132  (32.1%) | 23'631'920  (91.9%) |

**Coverage and bias**

Cumulative coverage was similar for the same side of each sample, but the two sides show a difference of ~10% for all coverage thresholds, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, with HM\_RS7\_G4\_2 showing some pattern for R2 reads in CHG and CHH context, both slightly overestimating % methylation in the second half of the reads. This slight bias is not a major concern and shouldn’t affect downstream analyses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_RS7\_G4\_1 | hal | 74.6% | 68.2% | 60.9% |
| lyr | 59.8% | 52.3% | 42.0% |
| HM\_RS7\_G4\_2 | hal | 75.6% | 70.1% | 59.9% |
| lyr | 60.9% | 53.7% | 42.7% |
| HM\_RS7\_G4\_3 | hal | 72.8% | 69.0% | 57.8% |
| lyr | 60.7% | 57.3% | 51.2% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_RS7\_G4\_1 | High | 9'712'647 | 40'452'708 | 19.36% | 50'165'355 (89%) |
| Low | 1'618'636 | 4'867'475 | 24.96% | 6'486'111 (11%) |
| HM\_RS7\_G4\_2 | High | 9'377'331 | 40'729'025 | 18.71% | 52'284'106 (92%) |
| Low | 1'579'820 | 4'965'290 | 24.14% | 6'545'110 (8%) |
| HM\_RS7\_G4\_3 | High | 9'723'019 | 38'045'227 | 20.35% | 47'768'246 (84%) |
| Low | 1'993'209 | 6'890'011 | 22.44% | 8'883'220 (16%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_RS7\_G4\_1 | High | 6'081'660 | 25'959'944 | 18.98% | 32'041'604 (87%) |
| Low | 1'170'677 | 3'728'549 | 23.90% | 4'899'226 (13%) |
| HM\_RS7\_G4\_2 | High | 5'927'795 | 26'578'131 | 18.24% | 32'505'926 (88%) |
| Low | 1'072'874 | 3'362'030 | 24.19% | 4'434'904 (12%) |
| HM\_RS7\_G4\_3 | High | 6'241'352 | 25'218'004 | 19.84% | 31'459'356 (85%) |
| Low | 1'317'661 | 4'163'813 | 24.04% | 5'481'474 (15%) |

For all samples, at least 84% of cytosines’ methylation state was called with high confidence on both sides. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 natural (ALK), HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | HM\_ALK\_G1\_1 | HM\_ALK\_G1\_2 | | HM\_ALK\_G1\_3 |
| MultiQC report | HM\_nat4v1\_ALK\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_ALK\_G1\_1 | 6.0% | 94.0% |
| HM\_ALK\_G1\_2 | 6.4% | 93.6% |
| HM\_ALK\_G1\_3 | 6.2% | 93.8% |

The conversion error rate was between around ~6%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed differences between sides of about 6%. Specifically, the lyr side showed lower mapping rates compared to the hal side. Deduplication rates were around 2-3%, a very low proportion of duplicated reads. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for this slight imbalance between subgenomes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_ALK\_G1\_1 | 68'516'347 | hal | 21’194’992  (30.9%) | 20’691’793  (97.6%) |
| lyr | 16’916’675  (24.7%) | 16’513’016  (97.6%) |
| HM\_ALK\_G1\_2 | 70'529'686 | hal | 21’166’116  (30.0%) | 20’637’954  (97.5%) |
| lyr | 16’944’614  (24.0%) | 16’519’425  (97.5%) |
| HM\_ALK\_G1\_3 | 60'808'951 | hal | 17’059’944  (28.1%) | 16’660’506  (97.7%) |
| lyr | 13’639’315  (22.4%) | 13’322’575  (97.7%) |

**Coverage and bias**

Cumulative coverage was similar for the same side of each sample, but the two sides show a difference of ~10% for all coverage thresholds, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, with HM\_RS7K\_G4\_2 showing some pattern for R2 reads in CHG and CHH context, both slightly overestimating % methylation in the second half of the reads. This slight bias is not a major concern and shouldn’t affect downstream analyses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_ALK\_G1\_1 | hal | 66.1% | 58.6% | 47.0% |
| lyr | 59.9% | 50.5% | 38.1% |
| HM\_ALK\_G1\_2 | hal | 66.1% | 58.6% | 46.8% |
| lyr | 59.8% | 50.4% | 37.8% |
| HM\_ALK\_G1\_3 | hal | 64.6% | 55.1% | 41.0% |
| lyr | 55.9% | 46.5% | 32.1% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_ALK\_G4\_1 | High | 10'100'995 | 39'766'864 | 20.26% | 49'867'859 (83%) |
| Low | 1'681'696 | 8'608'513 | 16.34% | 10'290'209 (17%) |
| HM\_ALK\_G4\_2 | High | 10'014'028 | 39'747'102 | 20.12% | 49'761'130 (83%) |
| Low | 1'712'445 | 8'684'493 | 16.47% | 10'396'938 (17%) |
| HM\_ALK\_G4\_3 | High | 10'019'292 | 39'353'668 | 20.29% | 49'372'960 (82%) |
| Low | 1'808'333 | 8'976'775 | 16.77% | 10'785'108 (18%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_ALK\_G4\_1 | High | 7'728'180 | 35'536'375 | 17.86% | 43'264'555 (79%) |
| Low | 1'649'958 | 10'098'749 | 14.04% | 11'748'707 (21%) |
| HM\_ALK\_G4\_2 | High | 7'688'608 | 35'618'474 | 17.75% | 43'307'082 (79%) |
| Low | 1'667'303 | 10'038'877 | 14.24% | 11'706'180 (21%) |
| HM\_ALK\_G4\_3 | High | 7'670'701 | 35'155'026 | 17.91% | 42'825'727 (78%) |
| Low | 1'759'848 | 10'427'687 | 14.44% | 12'187'535 (22%) |

For all samples, ~83% of cytosines’ methylation state was called with high confidence on the *halleri* side and ~79% had high confidence on the *lyrata* side. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 4 natural (ALK), HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 4 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | HM\_ALK\_G4\_1 | HM\_ALK\_G4\_2 | | HM\_ALK\_G4\_3 |
| MultiQC report | HM\_nat4v1\_ALK\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_ALK\_G4\_1 | 5.7% | 94.3% |
| HM\_ALK\_G4\_2 | 6.3% | 93.7% |
| HM\_ALK\_G4\_3 | 5.5% | 94.5% |

The conversion error rate was between 5-6%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples and generally similar in different sides, with the exception of HM\_ALK\_G4\_2 where the alignment difference was slightly bigger, with hal showing ~5% higher alignment rate. Alignment rate was lower than expected for all samples ~30% instead of an expected ~40%. Deduplication rates were around 4-5%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for the lower amount of aligned and deduplicated reads compared to other samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_ALK\_G4\_1 | 80'522'926 | hal | 25'367'453  (31.5%) | 23'084'625  (91.0%) |
| lyr | 20'240'923  (25.1%) | 18'415'354  (90.1%) |
| HM\_ALK\_G4\_2 | 104'369'803 | hal | 29'001'085  (27.8%) | 26'427'923  (91.1%) |
| lyr | 23'468'932  (22.5%) | 21'392'237  (91.2%) |
| HM\_ALK\_G4\_3 | 84'173'333 | hal | 25'218'963  (30.0%) | 22'910'321  (90.9%) |
| lyr | 19'794'968  (23.5%) | 17'988'431  (90.1%) |

**Coverage and bias**

Cumulative coverage was variable between samples and between different sides of the same sample. HM\_ALK\_G4\_2 had ~10% more % genome for 6X and 10X and ~5% for 15X compared to both HM\_ALK\_G4\_1 and 2. The two sides showed a difference of 5-10% for all coverage thresholds in all samples, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, showing no striking pattern in the % methylation along reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_ALK\_G4\_1 | hal | 66.5% | 59.3% | 48.2% |
| lyr | 60.2% | 51.3% | 39.2% |
| HM\_ALK\_G4\_2 | hal | 69.1% | 65.4% | 59.3% |
| lyr | 63.9% | 58.5% | 50.9% |
| HM\_ALK\_G4\_3 | hal | 65.2% | 56.5% | 44.7% |
| lyr | 58.7% | 48.3% | 35.8% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_ALK\_G4\_1 | High | 10'272'262 | 39'787'615 | 20.52% | 50'059'877 (83%) |
| Low | 1'563'129 | 8'535'062 | 15.48% | 10'098'191 (17%) |
| HM\_ALK\_G4\_2 | High | 10'447'364 | 40'154'319 | 20.65% | 50'601'683 (84%) |
| Low | 1'520'957 | 8'035'428 | 15.92% | 9'556'385 (16%) |
| HM\_ALK\_G4\_3 | High | 10'201'194 | 39'776'209 | 20.41% | 49'977'403 (82%) |
| Low | 1'568'143 | 8'612'522 | 15.40% | 10'180'665 (18%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_ALK\_G4\_1 | High | 7'907'215 | 35'666'670 | 18.15% | 43'573'885 (79%) |
| Low | 1'564'233 | 9'875'144 | 13.67% | 11'439'377 (21%) |
| HM\_ALK\_G4\_2 | High | 8'018'230 | 36'014'283 | 18.21% | 44'032'513 (80%) |
| Low | 1'518'005 | 9'462'744 | 13.82% | 10'980'749 (20%) |
| HM\_ALK\_G4\_3 | High | 7'867'329 | 35'596'731 | 18.10% | 43'464'060 (79%) |
| Low | 1'554'572 | 9'994'630 | 13.46% | 11'549'202 (21%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on both sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 natural (TKS), HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | HM\_TKS\_G1\_1 | HM\_TKS\_G1\_2 | | HM\_TKS\_G1\_3 |
| MultiQC report | HM\_nat5v1\_TKS\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_TKS\_G1\_1 | 4.6% | 95.4% |
| HM\_TKS\_G1\_2 | 4.8% | 95.2% |
| HM\_TKS\_G1\_3 | 4.2% | 95.8% |

The conversion error rate was ~4-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples and generally similar in different sides. Alignment rate was lower than expected for all samples: 20-30% instead of an expected ~40%. Deduplication rates were around 2-3%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for the lower amount of aligned and deduplicated reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_TKS\_G1\_1 | 66'904'897 | hal | 21’140’001  (31.6%) | 20’556’056  (97.2%) |
| lyr | 15’774’168  (23.6%) | 15’340’079  (97.3%) |
| HM\_TKS\_G1\_2 | 64'963'564 | hal | 20’907’561  (32.2%) | 20’381’442  (97.5%) |
| lyr | 15’646’621  (24.1%) | 15’251’911  (97.5%) |
| HM\_TKS\_G1\_3 | 79'258'067 | hal | 24’292’219  (30.6%) | 23’378’044  (96.2%) |
| lyr | 17’542’714  (22.1%) | 16’888’940  (96.3%) |

**Coverage and bias**

Cumulative coverage was similar between samples from the same side and between across sides. The two sides showed a difference of 10-15% for all coverage thresholds in all samples, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, showing no striking pattern in the % methylation along reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_TKS\_G1\_1 | hal | 66.9% | 60.4% | 49.3% |
| lyr | 56.2% | 47.3% | 35.2% |
| HM\_TKS\_G1\_2 | hal | 66.7% | 60.2% | 49.3% |
| lyr | 56.2% | 47.2% | 35.4% |
| HM\_TKS\_G1\_3 | hal | 66.9% | 59.9% | 48.4% |
| lyr | 56.3% | 46.7% | 34.7% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_TKS\_G1\_1 | High | 9'898'348 | 40'059'543 | 19.81% | 49'957'891 (83%) |
| Low | 1'856'887 | 8'343'290 | 18.20% | 10'200'177 (17%) |
| HM\_TKS\_G1\_2 | High | 9'947'258 | 39'844'083 | 19.98% | 49'791'341 (83%) |
| Low | 1'928'028 | 8'438'699 | 18.60% | 10'366'727 (17%) |
| HM\_TKS\_G1\_3 | High | 9'823'772 | 40'306'700 | 19.60% | 50'130'472 (83%) |
| Low | 1'792'194 | 8'235'402 | 17.87% | 10'027'596 (17%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_TKS\_G1\_1 | High | 7'188'541 | 34'654'257 | 17.18% | 41'842'798 (76%) |
| Low | 1'760'528 | 11'409'936 | 13.37% | 13'170'464 (24%) |
| HM\_TKS\_G1\_2 | High | 7'240'220 | 34'545'939 | 17.33% | 41'786'159 (76%) |
| Low | 1'841'070 | 11'386'033 | 13.92% | 13'227'103 (24%) |
| HM\_TKS\_G1\_3 | High | 7'173'628 | 34'844'435 | 17.07% | 42'018'063 (76%) |
| Low | 1'699'350 | 11'295'849 | 13.08% | 12'995'199 (24%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on the *halleri* side and ~76% had high confidence on the *lyrata* side. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 5 natural (TKS), HM conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 5 | | | |
| Conditions | High Mountain (HM) conditions: max T 22°C, winter 4°C for 8 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | HM\_TKS\_G5\_1 | HM\_TKS\_G5\_2 | | HM\_TKS\_G5\_3 |
| MultiQC report | HM\_nat5v1\_TKS\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| HM\_TKS\_G1\_1 | 3.7% | 96.3% |
| HM\_TKS\_G1\_2 | 3.8% | 96.2% |
| HM\_TKS\_G1\_3 | 4.5% | 95.5% |

The conversion error rate was ~4%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples and generally different in different sides. Alignment rate was lower than expected for all samples: 20-30% instead of an expected ~40%. Deduplication rates were around 3-4%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for the lower amount of aligned and deduplicated reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| HM\_TKS\_G5\_1 | 65'551'357 | hal | 21’075’639  (32.2%) | 20’403’165  (96.8%) |
| lyr | 15’430’883  (23.5%) | 14’937’894  (96.8%) |
| HM\_TKS\_G5\_2 | 68'305'559 | hal | 21’987’077  (32.2%) | 20’168’176  (96.3%) |
| lyr | 16’190’026  (23.7%) | 15’583’827  (96.3%) |
| HM\_TKS\_G5\_3 | 48'505'971 | hal | 15’506’983  (32.0%) | 15’147’941  (97.7%) |
| lyr | 11’898’923  (24.5%) | 11’620’212  (97.7%) |

**Coverage and bias**

Cumulative coverage was similar between samples from the same side and between across sides. The two sides showed a difference of 10-15% for all coverage thresholds in all samples, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, showing no striking pattern in the % methylation along reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| HM\_TKS\_G5\_1 | hal | 65.6% | 57.3% | 44.3% |
| lyr | 54.5% | 43.9% | 30.7% |
| HM\_TKS\_G5\_2 | hal | 66.3% | 58.9% | 46.5% |
| lyr | 55.5% | 45.7% | 32.7% |
| HM\_TKS\_G5\_3 | hal | 63.2% | 52.6% | 37.4% |
| lyr | 51.6% | 39.2% | 24.9% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_TKS\_G5\_1 | High | 9'943'902 | 39'936'377 | 19.94% | 49'880'279 (83%) |
| Low | 1'833'788 | 8'444'001 | 17.84% | 10'277'789 (17%) |
| HM\_TKS\_G5\_2 | High | 10'073'681 | 40'009'938 | 20.11% | 50'083'619 (83%) |
| Low | 1'804'194 | 8'270'255 | 17.91% | 10'074'449 (17%) |
| HM\_TKS\_G5\_3 | High | 9'995'371 | 38'987'713 | 20.41% | 48'983'084 (81%) |
| Low | 2'188'933 | 8'986'051 | 19.59% | 11'174'984 (19%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| HM\_TKS\_G5\_1 | High | 7'239'786 | 34'532'311 | 17.33% | 41'772'097 (76%) |
| Low | 1'751'419 | 11'489'746 | 13.23% | 13'241'165 (24%) |
| HM\_TKS\_G5\_2 | High | 7'336'783 | 34'701'302 | 17.45% | 42'038'085 (76%) |
| Low | 1'723'423 | 11'251'754 | 13.28% | 12'975'177 (24%) |
| HM\_TKS\_G5\_3 | High | 7'276'683 | 33'880'722 | 17.68% | 41'157'405 (75%) |
| Low | 2'071'283 | 11'784'574 | 14.95% | 13'855'857 (25%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on the *halleri* side and ~75% had high confidence on the *lyrata* side. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 synthetic, LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | | |
| Sample name (short) | LL\_RS7\_G1\_1 | LL\_RS7\_G1\_2 | | LL\_RS7\_G1\_3 |
| MultiQC report | LL\_tks1Vsyn1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_RS7 \_G1\_1 | 5.6% | 94.4% |
| LL\_RS7 \_G1\_2 | 4.8% | 95.2% |
| LL\_RS7 \_G1\_3 | 5.3% | 94.7% |

The conversion error rate was between ~5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting varied between samples and between sides. LL\_RS7K\_G4\_1 and 2 had a 5-9% difference in mapping rate between hal and lyr, with the latter having a lower proportion of uniquely mapped reads. LL\_RS7K \_G4\_1 had the lowest mapping rates for both sides, while LL\_RS7K \_G4\_2 had the highest for both. LL\_RS7K \_G4\_3 had similar rates between sides around 28%. Deduplication rates were around 7-8%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_RS7\_G1\_1 | 58'940'101 | hal | 22’489’844  (38.2%) | 21’877’395  (97.3%) |
| lyr | 19’594’152  (33.2%) | 19’059’901  (97.3%) |
| LL\_RS7\_G1\_2 | 58'281'178 | hal | 22’300’652  (38.3%) | 21’789’322  (97.7%) |
| lyr | 17’698’749  (30.4%) | 17’289’530  (97.7%) |
| LL\_RS7\_G1\_3 | 62'676'147 | hal | 23’805’085  (38.0%) | 23’045’791  (96.8%) |
| lyr | 19’937’282  (31.8%) | 19’300’324  (96.8%) |

The mapping rate after read sorting was similar between samples and generally different in different sides. Alignment rate was lower than expected for some samples: ~30% instead of an expected ~40%. Deduplication rates were around 3-4%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples, but some attention should be considered for the lower amount of aligned and deduplicated reads.

**Coverage and bias**

Cumulative coverage was similar between samples from the same side and between across sides. The two sides showed a difference of 10-15% for all coverage thresholds in all samples, with the hal side showing better coverage than lyr. This effect resulted from the previous alignment step, which showed a similar imbalance between subgenomes. M-bias plots were good for all samples, showing no striking pattern in the % methylation along reads.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_RS7\_G1\_1 | hal | 77.3% | 71.0% | 59.0% |
| lyr | 78.7% | 70.9% | 56.7% |
| LL\_RS7\_G1\_2 | hal | 78.6% | 73.7% | 62.9% |
| lyr | 79.2% | 66.7% | 50.3% |
| LL\_RS7\_G1\_3 | hal | 77.3% | 70.9% | 59.4% |
| lyr | 77.9% | 69.3% | 55.1% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_RS7\_G1\_1 | High | 11'940'587 | 42'942'610 | 21.76% | 54'883'197 (91%) |
| Low | 1'451'435 | 3'823'436 | 27.52% | 5'274'871 (9%) |
| LL\_RS7\_G1\_2 | High | 12'212'897 | 43'291'314 | 22.00% | 55'504'211 (92%) |
| Low | 1'274'256 | 3'379'601 | 23.38% | 4'653'857 (8%) |
| LL\_RS7\_G1\_3 | High | 11'833'158 | 43'542'413 | 21.37% | 55'375'571 (92%) |
| Low | 1'253'785 | 3'528'712 | 26.22% | 4'782'497 (8%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_RS7\_G1\_1 | High | 9'673'650 | 41'019'919 | 19.08% | 50'693'569 (92%) |
| Low | 1'126'417 | 3'193'276 | 26.08% | 4'319'693 (8%) |
| LL\_RS7\_G1\_2 | High | 9'734'448 | 41'149'370 | 19.13% | 50'883'818 (93%) |
| Low | 1'061'181 | 3'068'263 | 25.70% | 4'129'444 (7%) |
| LL\_RS7\_G1\_3 | High | 9'493'580 | 41'198'431 | 18.73% | 50'692'011 (92%) |
| Low | 991'778 | 3'329'473 | 22.95% | 4'321'251 (8%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on the *halleri* side and ~75% had high confidence on the *lyrata* side. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 4 synthetic, LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 4 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 09.2020 | | | 01.2019 |
| Sample name (short) | LL\_RS7K\_G4\_2 | LL\_RS7K\_G4\_3 | | LL\_RS7K\_G4\_1 |
| MultiQC report | LL\_syn4vpro1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, two out of three samples showed good mean quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion. Only LL\_RS7K\_G4\_1 showed low quality scores in the 115-135bp part and a skewed per sequence base content in the R1 reads, particularly for %T in the first 20-30bp of reads. Other parameters were good.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_RS7K \_G4\_1 | 2.6% | 97.4% |
| LL\_RS7K \_G4\_2 | 5.8% | 94.2% |
| LL\_RS7K \_G4\_3 | 8.3% | 91.7% |

The conversion error rate was between 3-8%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting varied between samples and between sides. LL\_RS7K\_G4\_1 and 2 had a 5-9% difference in mapping rate between hal and lyr, with the latter having a lower proportion of uniquely mapped reads. LL\_RS7K \_G4\_1 had the lowest mapping rates for both sides, while LL\_RS7K \_G4\_2 had the highest for both. LL\_RS7K \_G4\_3 had similar rates between sides around 28%. Deduplication rates were around 7-8%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_RS7K\_G4\_1 | 116'088'290 | hal | 22’245’266  (19.2%) | 20’706’326  (93.1%) |
| lyr | 16’024’903  (13.8%) | 14’921’434  (93.1%) |
| LL\_RS7K\_G4\_2 | 73'498'363 | hal | 28’038’789  (38.1%) | 25’965’987  (92.6%) |
| lyr | 21’293’399  (29.0%) | 19’717’568  (92.6%) |
| LL\_RS7K\_G4\_3 | 61'422'676 | hal | 24’268’195  (28.9%) | 22’706’805  (93.6%) |
| lyr | 16’694’009  (27.2%) | 15’626’714  (93.6%) |

**Coverage and bias**

Cumulative coverage was variable between samples and between different sides of the same sample. In all of the samples, the hal side showed 10% or more genome coverage for all selected thresholds compared to lyr side. Similarly to alignment, LL\_RS7K\_G4\_1 had the lowest coverage values and LL\_RS7K\_G4\_2 had the highest. M-bias plots were fine for two out of three samples, with LL\_RS7K\_G4\_1 showing some bias in the R1 reads, particularly for CHG and CHH context.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_RS7K\_G4\_1 | hal | 69.0% | 60.0% | 49.2% |
| lyr | 59.5% | 48.8 % | 37.1% |
| LL\_RS7K\_G4\_2 | hal | 77.3% | 71.6% | 61.7% |
| lyr | 68.9% | 61.4% | 49.7% |
| LL\_RS7K\_G4\_3 | hal | 74.2% | 67.8% | 57.3% |
| lyr | 60.8% | 49.7% | 36.3% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_RS7K\_G4\_1 | High | 11'067'449 | 41'639'888 | 21.00% | 52'707'337 (88%) |
| Low | 1'714'135 | 5'736'596 | 23.01% | 7'450'731 (12%) |
| LL\_RS7K\_G4\_2 | High | 11'562'222 | 43'571'252 | 20.97% | 55'133'474 (92%) |
| Low | 1'267'579 | 3'757'015 | 25.23% | 5'024'594 (8%) |
| LL\_RS7K\_G4\_3 | High | 11'280'209 | 42'304'876 | 21.05% | 53'585'085 (89%) |
| Low | 1'495'235 | 5'077'748 | 22.75% | 6'572'983 (11%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_RS7K\_G4\_1 | High | 8'345'238 | 35'174'743 | 19.18% | 43'519'981 (79%) |
| Low | 1'550'045 | 9'943'236 | 13.49% | 11'493'281 (21%) |
| LL\_RS7K\_G4\_2 | High | 8'762'094 | 37'135'507 | 19.09% | 45'897'601 (83%) |
| Low | 1'253'821 | 7'861'840 | 13.75% | 9'115'661 (17%) |
| LL\_RS7K\_G4\_3 | High | 8'295'000 | 34'722'825 | 19.28% | 43'017'825 (78%) |
| Low | 1'541'261 | 10'454'176 | 12.85% | 11'995'437 (22%) |

For all samples, ~90% of cytosines’ methylation state was called with high confidence on the *halleri* side and ~80% had high confidence on the *lyrata* side. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 natural (TKS), LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | LL\_TKS\_G1\_1 | LL\_TKS\_G1\_2 | | LL\_TKS\_G1\_3 |
| MultiQC report | LL\_nat1Vsyn1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_TKS\_G1\_1 | 4.5% | 95.5% |
| LL\_TKS\_G1\_2 | 4.9% | 95.1% |
| LL\_TKS\_G1\_3 | 4.2% | 95.8% |

The conversion error rate was between 4-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed a ~5% difference between sides, with hal showing higher values than lyr. Deduplication rates were around 3%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples and the final number of reads was similar across samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_TKS\_G1\_1 | 65'488'885 | hal | 21'382'192  (32.6%) | 20'798'262  (97.3%) |
| lyr | 15'977'836  (24.4%) | 15'542'924  (97.3%) |
| LL\_TKS\_G1\_2 | 63'982'362 | hal | 20'715'503  (32.4%) | 20'171'836  (97.4%) |
| lyr | 15'846'300  (24.8%) | 15'425'816  (97.4%) |
| LL\_TKS\_G1\_3 | 63'923'589 | hal | 21'169'347  (33.1%) | 20'580'446  (97.2%) |
| lyr | 15'864'588  (24.8%) | 15'421'985  (97.2%) |

**Coverage and bias**

Cumulative coverage was similar between samples, but different for the sides. Specifically, the lyr sides showed a consistent lower % genome of around 10-15% compared to the hal sides. These values are consistent with the result from the alignment step. M-bias plots were good for all three samples, showing no particular pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_TKS\_G1\_1 | hal | 67.3% | 61.3% | 50.9% |
| lyr | 56.9% | 48.4% | 36.8% |
| LL\_TKS\_G1\_2 | hal | 67.0% | 60.9% | 50.5% |
| lyr | 56.7% | 48.0% | 36.3% |
| LL\_TKS\_G1\_3 | hal | 66.3% | 59.0% | 47.6% |
| lyr | 55.6% | 46.0% | 33.9% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_TKS\_G1\_1 | High | 10'212'893 | 39'937'456 | 20.36% | 50'150'349 (83%) |
| Low | 1'834'032 | 8'173'687 | 18.33% | 10'007'719 (17%) |
| LL\_TKS\_G1\_2 | High | 10'409'949 | 39'645'542 | 20.80% | 50'055'491 (83%) |
| Low | 1'903'827 | 8'198'750 | 18.84% | 10'102'577 (17%) |
| LL\_TKS\_G1\_3 | High | 10'304'729 | 39'623'683 | 20.64% | 49'928'412 (83%) |
| Low | 1'873'622 | 8'356'034 | 18.32% | 10'229'656 (17%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_TKS\_G1\_1 | High | 7'408'748 | 34'665'291 | 17.61% | 42'074'039 (77%) |
| Low | 1'759'087 | 11'180'136 | 13.59% | 12'939'223 (23%) |
| LL\_TKS\_G1\_2 | High | 7'605'724 | 34'556'458 | 18.04% | 42'162'182 (77%) |
| Low | 1'873'099 | 10'977'981 | 14.58% | 12'851'080 (23%) |
| LL\_TKS\_G1\_3 | High | 7'483'094 | 34'347'800 | 17.89% | 41'830'894 (76%) |
| Low | 1'788'037 | 11'394'331 | 13.56% | 13'182'368 (24%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on both progenitors’ sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 5 natural (TKS), LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 06.2021 | | |  |
| Sample name (short) | LL\_TKS\_G5\_1 | LL\_TKS\_G5\_2 | | LL\_TKS\_G5\_3 |
| MultiQC report | LL\_tks5v1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_TKS\_G5\_1 | 5.5% | 94.5% |
| LL\_TKS\_G5\_2 | 5.1% | 94.9% |
| LL\_TKS\_G5\_3 | 5.1% | 94.9% |

The conversion error rate was between ~5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed a ~8% difference between sides, with hal showing higher values than lyr. Deduplication rates were around 3%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples and the final number of reads was similar across samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_TKS\_G5\_1 | 61'314'587 | hal | 19'925'883  (32.5%) | 19'401'318  (97.4%) |
| lyr | 14'854'156  (24.2%) | 14'463'986  (97.4%) |
| LL\_TKS\_G5\_2 | 92'015'167 | hal | 31'658'716  (34.4%) | 30'827'914  (97.4%) |
| lyr | 24'487'628  (26.6%) | 23'830'460  (97.3%) |
| LL\_TKS\_G5\_3 | 77'008'955 | hal | 24'483'937  (31.8%) | 23'769'596  (97.1%) |
| lyr | 18'247'523  (23.7%) | 17'718'115  (97.1%) |

**Coverage and bias**

Cumulative coverage was similar between samples, but different for the sides. Specifically, the lyr sides showed a consistent lower % genome of around 10-15% compared to the hal sides. These values are consistent with the result from the alignment step. M-bias plots were good for all three samples, showing no particular pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_TKS\_G1\_1 | hal | 67.5% | 58.3% | 46.5% |
| lyr | 57.3% | 45.1% | 32.7% |
| LL\_TKS\_G1\_2 | hal | 70.6% | 66.4% | 61.3% |
| lyr | 62.9% | 55.6% | 48.3% |
| LL\_TKS\_G1\_3 | hal | 69.1% | 62.9% | 54.0% |
| lyr | 59.9% | 50.4% | 39.9% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_TKS\_G5\_1 | High | 9'957'788 | 39'654'844 | 20.07% | 49'612'632 (82%) |
| Low | 1'938'801 | 8'606'635 | 18.39% | 10'545'436 (18%) |
| LL\_TKS\_G5\_2 | High | 10'617'627 | 41'093'834 | 20.53% | 51'711'461 (86%) |
| Low | 1'663'387 | 6'783'220 | 19.69% | 8'446'607 (12%) |
| LL\_TKS\_G5\_3 | High | 10'090'187 | 40'109'731 | 20.10% | 50'199'918 (83%) |
| Low | 1'790'316 | 8'167'834 | 17.98% | 9'958'150 (17%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_TKS\_G5\_1 | High | 7'215'495 | 34'260'269 | 17.40% | 41'475'764 (75%) |
| Low | 1'854'362 | 11'683'136 | 13.70% | 13'537'498 (25%) |
| LL\_TKS\_G5\_2 | High | 8'036'210 | 36'380'925 | 18.09% | 44'417'135 (75%) |
| Low | 1'812'888 | 8'783'239 | 17.11% | 10'596'127 (25%) |
| LL\_TKS\_G5\_3 | High | 7'360'192 | 34'835'844 | 17.44% | 42'196'036 (77%) |
| Low | 1'733'674 | 11'083'552 | 13.53% | 12'817'226 (23%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on both progenitors’ sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 1 natural (ALK), LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 1 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 01.2020 | | |  |
| Sample name (short) | LL\_ALK\_G1\_1 | LL\_ALK\_G1\_2 | | LL\_ALK\_G1\_3 |
| MultiQC report | LL\_nat4v1\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_ALK\_G1\_1 | 5.7% | 94.3% |
| LL\_ALK\_G1\_2 | 5.4% | 94.6% |
| LL\_ALK\_G1\_3 | 4.4% | 95.6% |

The conversion error rate was between ~5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed a ~5% difference between sides, with hal showing higher values than lyr. Deduplication rates were around 3%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples and the final number of reads was similar across samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_ALK\_G1\_1 | 93'692'564 | hal | 29'542'344  (31.9%) | 27'828'976  (94.2%) |
| lyr | 24'258'374  (26.2%) | 22'831'294  (94.1%) |
| LL\_ALK\_G1\_2 | 84'880'753 | hal | 28'561'679  (33.6%) | 26'815'497  (93.9%) |
| lyr | 23'553'565  (27.7%) | 22'089'187  (93.8%) |
| LL\_ALK\_G1\_3 | 76'125'831 | hal | 23'533'760  (31.1%) | 21'952'242  (93.3%) |
| lyr | 19'700'959  (26.0%) | 18'344'886  (93.1%) |

**Coverage and bias**

Cumulative coverage was similar between samples, but different for the sides. Specifically, the lyr sides showed a consistent lower % genome of around 10% compared to the hal sides. These values are consistent with the result from the alignment step. M-bias plots were good for all three samples, showing no particular pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_ALK\_G1\_1 | hal | 70.2% | 65.6% | 59.8% |
| lyr | 65.7% | 59.0% | 51.6% |
| LL\_ALK\_G1\_2 | hal | 69.9% | 65.1% | 58.9% |
| lyr | 65.2% | 58.3% | 50.4% |
| LL\_ALK\_G1\_3 | hal | 68.8% | 63.2% | 55.2% |
| lyr | 63.7% | 55.6% | 45.9% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_ALK\_G1\_1 | High | 10'883'687 | 40'165'766 | 21.32% | 51'049'453 (85%) |
| Low | 1'566'890 | 7'541'725 | 17.20% | 9'108'615 (15%) |
| LL\_ALK\_G1\_2 | High | 10'916'911 | 39'902'716 | 21.48% | 50'819'627 (85%) |
| Low | 1'601'650 | 7'736'791 | 17.15% | 9'338'441 (15%) |
| LL\_ALK\_G1\_3 | High | 10'885'179 | 39'501'638 | 21.60% | 50'386'817 (84%) |
| Low | 1'684'245 | 8'087'006 | 17.24% | 9'771'251 (16%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_ALK\_G1\_1 | High | 8'489'903 | 36'280'953 | 18.96% | 44'770'856 (81%) |
| Low | 1'597'672 | 8'644'734 | 15.60% | 10'242'406 (19%) |
| LL\_ALK\_G1\_2 | High | 8'476'956 | 35'999'346 | 19.06% | 44'476'302 (81%) |
| Low | 1'627'322 | 8'909'638 | 15.44% | 10'536'960 (19%) |
| LL\_ALK\_G1\_3 | High | 8'416'159 | 35'521'782 | 19.15% | 43'937'941 (80%) |
| Low | 1'711'856 | 9'363'465 | 15.46% | 11'075'321 (20%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on both progenitors’ sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

# ***Arabidopsis kamchatica* generation 4 natural (ALK), LL conditions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | *Arabidopsis kamchatica* | | | |
| Generation | Generation 4 | | | |
| Conditions | Low Land (LL) conditions: max T 26°C, winter 4°C for 4 weeks | | | |
| Sequencing platform | NovaSeq 6000 | | | |
| Date | 01.2020 | | |  |
| Sample name (short) | LL\_ALK\_G4\_1 | LL\_ALK\_G4\_2 | | LL\_ALK\_G4\_3 |
| MultiQC report | HMvLL\_nat4v4\_multiqc\_report.html | |  |  |

**Trimming and quality check**

After a preliminary quality check, all samples were trimmed 10bp in the 3’ end and 5bp 5’, and all adapter sequences were removed. After trimming, all of the samples showed good quality scores, per sequence quality scores, per base sequence content, per base N content, duplication levels, overrepresented sequences and no adapters. Both warnings in per base sequence content and sequence GC content are expected because of the effect of bisulfite conversion.

**In silico conversion check**

|  |  |  |
| --- | --- | --- |
|  | Conversion error rate | Conversion efficiency |
| LL\_ALK \_G4\_1 | 6.7% | 95.3% |
| LL\_ALK \_G4\_2 | 5.4% | 94.6% |
| LL\_ALK \_G4\_3 | 5.4% | 94.6% |

The conversion error rate was between 4-5%. The conversion process affected the library preparation step, because a stronger treatment would cause DNA to break more and lead to shorter fragments. These values represent a good trade-off between bisulphite conversion and library preparation.

**Alignment and deduplication**

The mapping rate after read sorting was similar between samples, but showed a ~5% difference between sides, with hal showing higher values than lyr. Deduplication rates were around 7-8%, meaning that limited reads were lost in the process. Overall, the outcome for alignment and deduplication was good for all of the samples and the final number of reads was related to the total number of initial reads, which was highest for LL\_ALK\_G4\_1, followed by 2 and then 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total reads (pairs) | Side | Uniquely mapped reads (%) | Deduplicated reads (%) |
| LL\_ALK\_G4\_1 | 101'409'930 | hal | 34’041’742  (33.6%) | 31’613’059  (92.9%) |
| lyr | 28’575’350  (28.2%) | 26’498’558  (92.7%) |
| LL\_ALK\_G4\_2 | 83'728'344 | hal | 26’490’841  (31.6%) | 24’508’203  (92.5%) |
| lyr | 21’756’947  (26.0%) | 20’099’070  (92.4%) |
| LL\_ALK\_G4\_3 | 64'457'303 | hal | 21’213’488  (32.9%) | 19’643’043  (92.6%) |
| lyr | 17’430’306  (27.0%) | 16’128’334  (92.5%) |

**Coverage and bias**

Cumulative coverage was similar between samples, but different for the sides. Specifically, the lyr sides showed a consistent lower % genome of around 5-10% compared to the hal sides. These values are consistent with the result from the alignment step. M-bias plots were good for all three samples, showing no particular pattern.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Side | % genome with at least 6X coverage | 10X coverage | 15X coverage |
| LL\_ALK\_G4\_1 | hal | 69.7% | 66.7% | 62.6% |
| lyr | 65.0% | 60.5% | 54.8% |
| LL\_ALK\_G4\_2 | hal | 68.2% | 64.0% | 57.2% |
| lyr | 62.2% | 55.7% | 46.6% |
| LL\_ALK\_G4\_3 | hal | 65.9% | 58.7% | 47.1% |
| lyr | 59.6% | 50.6% | 38.0% |

**Confidence of methylation calls**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **halleri** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_ALK\_G4\_1 | High | 11'180'314 | 39'831'079 | 21.92% | 51'011'393 (85%) |
| Low | 1'600'905 | 7'545'770 | 17.50% | 9'146'675 (15%) |
| LL\_ALK\_G4\_2 | High | 10'724'420 | 39'639'288 | 21.29% | 50'363'708 (84%) |
| Low | 1'666'378 | 8'127'982 | 17.01% | 9'794'360 (16%) |
| LL\_ALK\_G4\_3 | High | 10'431'156 | 39'284'648 | 20.98% | 49'715'804 (83%) |
| Low | 1'783'790 | 8'658'474 | 17.08% | 10'442'264 (17%) |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **lyrata** | Confidence | Methylated | Not methylated | % | Total (%) |
| LL\_ALK\_G4\_1 | High | 8'716'899 | 36'057'430 | 19.47% | 44'774'329 (81%) |
| Low | 1'635'839 | 8'603'094 | 15.98% | 10'238'933 (19%) |
| LL\_ALK\_G4\_2 | High | 8'277'681 | 35'626'214 | 18.85% | 43'903'895 (80%) |
| Low | 1'698'536 | 9'410'831 | 15.29% | 11'109'367 (20%) |
| LL\_ALK\_G4\_3 | High | 8'046'664 | 35'273'465 | 18.58% | 43'320'129 (79%) |
| Low | 1'768'612 | 9'924'521 | 15.13% | 11'693'133 (21%) |

For all samples, ~80% of cytosines’ methylation state was called with high confidence on both progenitors’ sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.