Sample quality report

Last update: 29.11.2022

Table of Contents

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, mild 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	COLD_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. 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 bisulfite 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 60% for the two samples from 2020 and at about half (~32%) for the sample from 2019, but with higher number of total reads, leading to a similar 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	25'164'186 (65.0%)	23'788'085 (94.5%)
HM_lyr_G1_2	27'272'244	17'167'399 (62.9%)	16'383'905 (95.4%)
HM_lyr_G1_3	62'028'002	19'827'484 (32.0%)	18'019'899 (90.9%)

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	86.3%	79.4%	67.1%
HM_lyr_G1_2	81.3%	64.2%	46.1%
HM_lyr_G1_3	77.1%	63.2%	50.7%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
HM_lyr_G1_1	High	11'035'222	34'279'938	24.35%	45'315'160 (92%)
	Low	1'265'084	2'867'939	30.61%	4'133'023 (8%)
HM_lyr_G1_2	High	10'813'664	33'696'892	24.29%	44'510'556 (90%)
	Low	1'489'614	3'448'013	30.17%	4'937'627 (10%)
HM_lyr_G1_3	High	10'916'152	32'251'977	25.29%	43'168'129 (87%)
	Low	1'921'038	4'359'016	30.59%	6'280'054 (13%)

For all three samples, around 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, mild 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	COLD_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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after alignment was very similar between samples, around ~60%. 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	26'257'544	15'905'574 (60.6%)	15'171'931 (95.4%)
HM_lyr_G4_2	30'030'523	18'462'556 (61.5%)	17'403'782 (94.3%)
HM_lyr_G4_3	27'600'085	16'945'000 (61.4%)	16'104'588 (95.0%)

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	78.2%	64.5%	46.7%
HM_lyr_G4_2	80.5%	69.0%	52.3%
HM_lyr_G4_3	79.9%	66.9%	48.2%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
HM_lyr_G4_1	High	10'493'283	33'624'932	23.78%	44'118'215 (89%)
	Low	1'633'726	3'696'242	30.65%	5'329'968 (11%)
HM_lyr_G4_2	High	10'882'788	33'513'234	24.51%	44'396'022 (90%)
	Low	1'554'538	3'497'623	30.77%	5'052'161 (10%)
HM_lyr_G4_3	High	10'876'303	33'559'050	24.48%	44'435'353 (90%)
	Low	1'537'754	3'475'076	30.68%	5'012'830 (10%)

For all three samples, around 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 1, stress 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	HOT_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
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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after alignment was around 60% 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	26'486'461	16'655'030 (62.9%)	15'991'138 (96.0%)
LL_lyr_G1_2	24'644'478	14'307'251 (58.1%)	13'620'349 (95.2%)
LL_lyr_G1_3	32'273'314	20'096'380 (62.3%)	19'208'186 (95.6%)

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	83.4%	77.4%	63.8%
LL_lyr_G1_2	81.8%	73.4%	54.0%
LL_lyr_G1_3	83.7%	75.2%	60.7%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
LL_lyr_G1_1	High	11'456'115	37'155'316	23.57%	48'611'431 (91%)
	Low	1'720'905	3'199'178	34.98%	4'920'083 (9%)
LL_lyr_G1_2	High	11'547'185	36'641'266	23.96%	48'188'451 (90%)
	Low	1'940'498	3'402'565	36.32%	5'343'063 (10%)
LL_lyr_G1_3	High	11'368'750	37'883'052	23.08%	49'251'802 (92%)
	Low	1'336'535	2'943'177	31.23%	4'279'712 (8%)

For all three samples, around 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, stress 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		06.2021
Sample name (short)	LL_lyr_G4_1	LL_lyr_G4_2	LL_lyr_G4_3
MultiQC report	HOT_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
LL_lyr_G4_1	6.2%	93.8%
LL_lyr_G4_2	5.5%	94.5%
LL_lyr_G4_3	8.2%	91.8%

The conversion error rate included 5-8% 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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after alignment was slightly over 55% for all samples, leading to an amount of uniquely mapped reads in a similar range. Deduplication rates were around 3-8%, 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	43'517'401	25'034'019 (57.5%)	23'183'138 (92.6%)
LL_lyr_G4_2	35'038'632	19'437'090 (55.5%)	18'160'290 (93.4%)
LL_lyr_G4_3	32'673'279	19'607'067 (60.0%)	19'020'283 (97.0%)

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_G4_1	84.6%	79.3%	69.3%
LL_lyr_G4_2	83.0%	76.1%	62.6%
LL_lyr_G4_3	84.1%	80.0%	70.6%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
LL_lyr_G4_1	High	12'141'353	37'197'822	24.61%	49'339'175 (92%)
	Low	1'313'822	2'878'517	31.34%	4'192'339 (8%)
LL_lyr_G4_2	High	11'992'510	37'010'586	24.47%	49'003'096 (92%)
	Low	1'470'405	3'058'013	32.47%	4'528'418 (8%)
LL_lyr_G4_3	High	11'638'535	37'145'663	23.86%	48'784'198 (91%)
	Low	1'645'955	3'101'361	34.67%	4'747'316 (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 halleri generation 1, mild 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	COLD_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. 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 bisulfite 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 4-5%, 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	17'774'109 (58.2%)	16'951'407 (95.4%)
HM_hal_G1_2	24'216'644	12'666'011 (52.3%)	12'151'470 (95.9%)
HM_hal_G1_3	40'168'222	11'347'741 (28.3%)	10'485'439 (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	79.1%	63.0%	47.7%
HM_hal_G1_2	80.2%	66.5%	44.1%
HM_hal_G1_3	63.8%	43.7%	28.9%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
HM_hal_G1_1	High	16'091'000	46'656'586	25.64%	62'747'586 (87%)
	Low	2'664'623	6'948'878	27.72%	9'613'501 (13%)
HM_hal_G1_2	High	16'320'867	45'543'575	26.38%	61'864'442 (86%)
	Low	3'273'648	7'222'997	31.19%	10'496'645 (14%)
HM_hal_G1_3	High	14'651'700	44'190'349	24.90%	58'842'049 (81%)
	Low	3'847'675	9'671'363	28.46%	13'519'038 (19%)

For all three samples, at least ~80% 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, mild 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	COLD_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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after alignment was very similar between samples, between 54-56%. 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	32'273'434	17'419'928 (54.0%)	16'715'108 (96.0%)
HM_hal_G4_2	27'628'204	15'604'484 (56.5%)	14'990'730 (96.1%)
HM_hal_G4_3	25'439'038	14'427'307 (56.7%)	13'787'732 (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_G4_1	79.2%	68.3%	51.2%
HM_hal_G4_2	75.9%	61.0%	42.6%
HM_hal_G4_3	70.8%	54.5%	38.0%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
HM_hal_G4_1	High	15'517'929	47'316'523	24.70%	62'834'452 (87%)
	Low	2'695'664	6'830'971	28.30%	9'526'635 (13%)
HM_hal_G4_2	High	15'324'274	47'387'505	24.44%	62'711'779 (87%)
	Low	2'684'187	6'965'121	27.82%	9'649'308 (13%)
HM_hal_G4_3	High	15'194'942	46'801'042	24.51%	61'995'984 (86%)
	Low	2'875'057	7'490'046	27.74%	10'365'103 (14%)

For all three samples, at least 86% 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, stress 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	HOT_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. 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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after alignment was around 59% 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	29'335'494	17'229'241 (58.7%)	16'581'928 (96.2%)
LL_hal_G1_2	30'973'234	18'298'386 (59.1%)	17'385'459 (95.0%)
LL_hal_G1_3	29'891'994	17'503'872 (58.6%)	16'712'139 (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	80.3%	70.7%	54.4%
LL_hal_G1_2	78.4%	67.2%	50.9%
LL_hal_G1_3	77.5%	64.4%	47.3%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
LL_hal_G1_1	High	16'790'316	46'767'400	26.42%	63'557'716 (88%)
	Low	2'553'126	6'044'342	29.70%	8'597'468 (12%)
LL_hal_G1_2	High	16'604'594	46'660'921	26.25%	63'265'515 (88%)
	Low	2'596'448	6'293'221	29.21%	8'889'669 (12%)
LL_hal_G1_3	High	16'511'274	46'943'424	26.02%	63'454'698 (88%)
	Low	2'530'116	6'170'370	29.08%	8'700'486 (12%)

For all three samples, 88% 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, stress 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	HOT_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. 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 bisulfite 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	36'563'014	18'489'808 (50.6%)	17'431'971 (94.3%)
LL_hal_G4_2	31'573'514	16'461'431 (52.1%)	15'493'324 (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	80.4%	74.2%	61.3%
LL_hal_G4_2	78.4%	67.7%	50.6%

Confidence of methylation calls

	Confidence	Methylated	Not methylated	%	Total (%)
LL_hal_G4_1	High	16'946'473	45'976'118	26.93%	62'922'591 (87%)
	Low	2'761'388	6'471'205	29.91%	9'232'593 (13%)
LL_hal_G4_2	High	16'629'330	46'170'339	26.48%	62'799'669 (87%)
	Low	2'784'100	6'571'415	29.76%	9'355'515 (13%)

For all three samples, ~87% 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, mild 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	COLD_syn4v1_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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate 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 ~30% less uniquely aligned reads. HM_RS7K_G1_2 and 3 were quite similar in their 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	19'770'856 (17.7%)	18'141'685 (91.8%)
		lyr	21'122'840 (18.9%)	19'301'319 (91.4%)
HM_RS7K_G1_2	82'499'019	hal	30'393'247 (36.8%)	28'452'353 (93.6%)
		lyr	27'154'920	25'312'008

			(32.9%)	(93.2%)
HM_RS7K_G1_3	73'276'049	hal	25'932'261 (35.4%)	24'017'557 (92.6%)
		lyr	28'536'092 (38.9%)	26'376'390 (92.4%)

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	72.4%	56.0%	42.0%
	lyr	75.7%	61.5%	48.3%
HM_RS7K_G1_2	hal	81.7%	75.3%	67.0%
	lyr	79.0%	69.2%	56.5%
HM_RS7K_G1_3	hal	80.1%	69.9%	57.1%
	lyr	84.1%	77.0%	66.7%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_RS7K_G1_1	High	15'548'309	44'561'671	25.87%	60'109'980 (83%)
	Low	3'341'740	8'909'367	27.28%	12'251'107 (17%)
HM_RS7K_G1_2	High	16'332'132	46'289'582	26.08%	62'621'714 (87%)
	Low	2'625'603	7'113'770	26.96%	9'739'373 (13%)
HM_RS7K_G1_3	High	15'825'944	45'987'803	25.60%	61'813'747 (85%)
	Low	2'918'677	7'628'663	27.67%	10'547'340 (15%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_RS7K_G1_1	High	10'392'931	32'389'546	24.29%	42'782'477 (87%)
	Low	1'789'431	4'876'275	26.85%	6'665'706 (13%)
HM_RS7K_G1_2	High	10'757'467	33'514'799	24.30%	44'272'266 (90%)

	Low	1'518'761	3'657'156	29.34%	5'175'917 (10%)
HM_RS7K_G1_3	High	10'747'988	33'937'192	24.05%	44'685'180 (90%)
	Low	1'406'931	3'356'072	29.54%	4'763'003 (10%)

For all three samples, ~85-90% 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, mild 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	COLD_syn4v1_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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate was similar between samples, around 30-40%, but showed slight differences between sides between 5-8%. Specifically, the lyr side showed lower mapping rates compared to the hal side. Deduplication rates were around 8-10%, 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	33'164'343 (44.5%)	29'692'331 (89.5%)
		lyr	27'103'051 (36.4%)	24'131'047 (89.0%)
HM_RS7_G4_2	75'226'540	hal	28'474'385 (37.9%)	26'407'858 (92.7%)

		lyr	23'025'098 (30.6%)	21'252'722 (92.3%)
HM_RS7_G4_3	80'065'803	hal	34'977'278 (43.7%)	32'111'375 (91.8%)
		lyr	30'749'473 (38.4%)	28'061'355 (91.3%)

Coverage and bias

Cumulative coverage was similar for the same side of each sample, but the two sides show a difference of ~10-15% 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.

	Side	% genome with at least 6X coverage	10X coverage	15X coverage
HM_RS7_G4_1	hal	78.0%	68.5%	57.5%
	lyr	63.4%	53.5%	43.0%
HM_RS7_G4_2	hal	79.3%	70.6%	64.1%
	lyr	64.8%	55.0%	43.6%
HM_RS7_G4_3	hal	76.5%	70.0%	62.2%
	lyr	64.1%	58.9%	52.2%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_RS7_G4_1	High	14'586'603	47'911'870	23.34%	62'498'473 (86%)
	Low	2'491'913	7'370'701	25.27%	9'862'614 (14%)
HM_RS7_G4_2	High	14'816'200	47'783'837	23.67%	62'600'037 (87%)
	Low	2'434'081	7'326'969	24.94%	9'761'050 (13%)
HM_RS7_G4_3	High	14'733'661	45'365'575	24.52%	60'099'236 (83%)
	Low	2'748'011	9'513'840	22.41%	12'261'851 (17%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_RS7_G4_1	High	9'487'923	34'548'727	21.55%	44'036'650 (89%)
	Low	1'429'038	3'982'495	26.41%	5'411'533 (11%)

HM_RS7_G4_2	High	9'668'556	34'733'376	21.78%	44'401'932 (90%)
	Low	1'369'618	3'676'633	27.14%	5'046'251 (10%)
HM_RS7_G4_3	High	9'745'167	33'147'567	22.72%	42'892'734 (87%)
	Low	1'526'043	5'029'406	23.28%	6'555'449 (13%)

For all samples, at least 83% 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), mild 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	COLD_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 bisulfite 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: 28-33% 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.

	Total reads (pairs)	Side	Uniquely mapped reads (%)	Deduplicated reads (%)
HM_ALK_G1_1	68'516'148	hal	22'310'955 (32.6%)	21'725'451 (97.4%)
		lyr	20'592'291 (30.1%)	19'928'186 (96.8%)
HM_ALK_G1_2	70'529'511	hal	22'398'625 (31.8%)	21'769'844 (97.2%)
		lyr	20'935'738 (29.7%)	20'197'748 (96.5%)
HM_ALK_G1_3	60'808'797	hal	18'130'978 (29.8%)	17'654'920 (97.4%)

		lyr	17'059'781 (28.1%)	16'510'641 (96.8%)

Coverage and bias

Cumulative coverage was similar for the same side of each sample, but the two sides show a difference of ~5% 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	66.8%	62.4%	56.4%
	lyr	61.8%	56.1%	48.8%
HM_ALK_G1_2	hal	66.3%	61.7%	55.3%
	lyr	61.2%	55.3%	47.6%
HM_ALK_G1_3	hal	64.9%	59.7%	51.7%
	lyr	59.4%	52.6%	43.2%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_ALK_G1_1	High	12'167'638	40'993'838	22.89%	53'161'476 (73%)
	Low	3'951'590	15'248'021	20.58%	19'199'611 (27%)
HM_ALK_G1_2	High	12'040'246	40'972'035	22.71%	53'012'281 (73%)
	Low	3'986'586	15'362'220	20.60%	19'348'806 (27%)
HM_ALK_G1_3	High	11'980'007	40'480'180	22.84%	52'460'187 (72%)
	Low	4'137'794	15'763'106	20.79%	19'900'900 (28%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_ALK_G1_1	High	7'600'960	27'842'611	21.45%	35'443'571 (72%)
	Low	2'732'830	11'271'782	19.51%	14'004'612 (28%)
HM_ALK_G1_2	High	7'584'384	27'894'344	21.38%	35'478'728 (72%)
	Low	2'724'019	11'245'436	19.50%	13'969'455 (28%)

HM_ALK_G1_3	High	7'516'964	27'470'801	21.48%	34'987'765 (71%)
	Low	2'897'222	11'563'196	20.04%	14'460'418 (29%)

For all samples, >70% of cytosines' methylation state was called with high confidence on both parental sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

Arabidopsis kamchatica generation 4 natural (ALK), mild 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	COLD_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 bisulfite conversion and library preparation.

Alignment and deduplication

Alignment rate was lower than expected for all samples ~30% instead of an expected ~40%. Deduplication rates were around 10%, 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 (%)
HM_ALK_G4_1	80'522'630	hal	26'365'437 (32.7%)	23'944'934 (90.8%)
		lyr	24'222'904 (30.1%)	21'852'065 (90.2%)
HM_ALK_G4_2	104'369'606	hal	30'495'728 (29.2%)	27'685'177 (90.8%)
		lyr	29'249'552 (28.0%)	26'331'314 (90.0%)
HM_ALK_G4_3	84'173'333	hal	26'404'022 (31.4%)	23'920'309 (90.6%)

		lyr	23'977'299 (28.5%)	21'537'030 (89.8%)

Coverage and bias

Cumulative coverage was variable between samples and between different sides of the same sample. 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	63.8%	56.4%	45.7%
	lyr	57.7%	48.9%	37.5%
HM_ALK_G4_2	hal	66.8%	62.3%	56.0%
	lyr	61.4%	55.7%	48.2%
HM_ALK_G4_3	hal	62.8%	54.0%	42.6%
	lyr	56.4%	46.3%	34.6%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_ALK_G4_1	High	12'471'584	41'016'704	23.32%	53'488'288 (74%)
	Low	3'799'230	15'073'569	20.13%	18'872'799 (26%)
HM_ALK_G4_2	High	12'679'273	41'459'844	23.42%	54'139'117 (75%)
	Low	3'693'517	14'528'453	20.27%	18'221'970 (25%)
HM_ALK_G4_3	High	12'425'906	41'025'775	23.25%	53'451'681 (74%)
	Low	3'824'386	15'085'020	20.22%	18'909'406 (26%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_ALK_G4_1	High	7'833'009	27'929'010	21.90%	35'762'019 (72%)
	Low	2'639'796	11'046'368	19.29%	13'686'164 (28%)
HM_ALK_G4_2	High	7'973'004	28'334'694	21.96%	36'307'698 (73%)
	Low	2'539'033	10'601'452	19.32%	13'140'485 (27%)

HM_ALK_G4_3	High	7'797'670	27'846'532	21.88%	35'644'202 (72%)
	Low	2'627'909	11'176'072	19.04%	13'803'981 (28%)

For all samples, >70% 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), mild 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	COLD_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 bisulfite 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: 25-30% 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	66'904'764	hal	21'232'128 (31.7%)	20'603'479 (97.0%)
		lyr	18'465'086 (27.6%)	17'830'970 (96.6%)
HM_TKS_G1_2	64'963'365	hal	21'210'788 (32.7%)	20'636'921 (97.3%)
		lyr	18'435'982 (28.4%)	17'860'117 (96.9%)
HM_TKS_G1_3	79'257'934	hal	24'962'310 (31.5%)	23'960'359 (96.0%)

		lyr	20'842'161 (26.3%)	19'873'623 (95.4%)

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 slight 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	64.2%	57.3%	46.6%
	lyr	53.3%	44.5%	33.2%
HM_TKS_G1_2	hal	63.9%	56.9%	46.4%
	lyr	53.2%	44.5%	33.4%
HM_TKS_G1_3	hal	64.4%	57.0%	46.0%
	lyr	53.6%	44.4%	33.0%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_TKS_G1_1	High	11'962'044	41'430'825	22.40%	53'392'869 (74%)
	Low	3'833'787	15'134'431	20.21%	18'968'218 (26%)
HM_TKS_G1_2	High	11'938'280	41'153'873	22.49%	53'092'153 (73%)
	Low	3'947'141	15'321'793	20.48%	19'268'934 (27%)
HM_TKS_G1_3	High	11'934'564	41'808'629	22.21%	53'743'193 (74%)
	Low	3'705'505	14'912'389	19.90%	18'617'894 (26%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_TKS_G1_1	High	6'909'747	27'110'569	20.31%	34'020'316 (69%)
	Low	2'899'376	12'528'491	18.79%	15'427'867 (31%)
HM_TKS_G1_2	High	6'957'675	27'024'089	20.48%	33'981'764 (69%)
	Low	2'939'020	12'527'399	19.00%	15'466'419 (31%)
HM_TKS_G1_3	High	6'974'171	27'316'728	20.34%	34'290'899 (69%)

	Low	2'795'328	12'361'956	18.44%	15'157'284 (31%)

For all samples, ~70% of cytosines' methylation state was called with high confidence on the *halleri* side and ~70% 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), mild 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	COLD_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_G5_1	3.7%	96.3%
HM_TKS_G5_2	3.8%	96.2%
HM_TKS_G5_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 bisulfite 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: 25-30% 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	65'551'182	hal	21'229'851 (32.4%)	20'518'428 (96.7%)
		lyr	17'730'957 (27.0%)	17'073'111 (96.3%)
HM_TKS_G5_2	68'305'366	hal	22'171'042 (32.5%)	21'318'992 (96.2%)
		lyr	18'714'327 (27.4%)	17'916'374 (95.7%)
HM_TKS_G5_3	48'505'794	hal	15'523'751	15'158'625

			(32.0%)	(97.7%)
		lyr	13'685'813 (28.2%)	13'336'298 (97.5%)

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	62.9%	54.4%	42.0%
	lyr	51.8%	41.5%	29.2%
HM_TKS_G5_2	hal	63.6%	55.9%	44.0%
	lyr	52.7%	43.2%	31.0%
HM_TKS_G5_3	hal	59.7%	49.2%	34.8%
	lyr	48.5%	36.7%	23.4%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
HM_TKS_G5_1	High	12'000'316	41'201'472	22.56%	53'201'788 (74%)
	Low	3'861'066	15'298'233	20.15%	19'159'299 (26%)
HM_TKS_G5_2	High	12'211'743	41'293'290	22.82%	53'505'033 (74%)
	Low	3'826'785	15'029'269	20.29%	18'856'054 (26%)
HM_TKS_G5_3	High	11'815'640	40'013'060	22.80%	51'828'700 (72%)
	Low	4'325'924	16'206'463	21.07%	20'532'387 (28%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
HM_TKS_G5_1	High	6'985'072	26'932'670	20.59%	33'917'742 (69%)
	Low	2'851'385	12'679'056	18.36%	15'530'441 (31%)
HM_TKS_G5_2	High	7'122'514	27'114'722	20.80%	34'237'236 (69%)
	Low	2'837'125	12'373'822	18.65%	15'210'947 (31%)

HM_TKS_G5_3	High	6'891'429	26'286'551	20.77%	33'177'980 (67%)
	Low	3'189'804	13'080'399	19.61%	16'270'203 (33%)

For all samples, ~70% of cytosines' methylation state was called with high confidence on both parental sides. Low confidence cytosines show a lower methylation proportion compared to the high confidence cytosines.

Arabidopsis kamchatica generation 1 synthetic, stress 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	HOT_syn4v1_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 bisulfite 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 as 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 (%)
LL_RS7_G1_1	58'939'904	hal	22'931'356 (38.9%)	22'309'215 (97.3%)
		lyr	22'642'266 (38.4%)	21'972'340 (97.0%)
LL_RS7_G1_2	58'280'818	hal	23'137'144 (39.7%)	22'607'376 (97.7%)
		lyr	20'697'630 (35.5%)	20'178'842 (97.5%)
LL_RS7_G1_3	62'675'984	hal	24'980'273 (39.9%)	24'176'079 (96.8%)

		lyr	23'662'981 (37.8%)	22'835'831 (96.5%)

Coverage and bias

Cumulative coverage was similar between samples from the same side and between across sides. The two sides showed a difference of 3-10% for all coverage thresholds in all samples. 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	69.0%	56.0%	42.0%
	lyr	72.8%	61.5%	48.3%
LL_RS7_G1_2	hal	80.5%	75.3%	67.0%
	lyr	77.4%	69.2%	56.5%
LL_RS7_G1_3	hal	78.3%	69.9%	57.1%
	lyr	82.8%	77.0%	66.7%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_RS7_G1_1	High	15'924'483	46'451'705	25.53%	62'376'188 (86%)
	Low	2'789'700	6'989'296	28.53%	9'778'996 (14%)
LL_RS7_G1_2	High	16'421'883	47'145'099	25.83%	63'566'982 (88%)
	Low	2'452'691	6'135'511	28.56%	8'588'202 (12%)
LL_RS7_G1_3	High	15'874'243	47'576'512	25.02%	63'450'755 (88%)
	Low	2'403'686	6'300'743	27.61%	8'704'429 (12%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
LL_RS7_G1_1	High	11'216'543	37'498'510	23.02%	48'715'053 (91%)
	Low	1'539'235	3'277'226	31.96%	4'816'461 (9%)
LL_RS7_G1_2	High	11'263'271	37'627'122	23.04%	48'890'393 (91%)
	Low	1'478'084	3'163'037	31.85%	4'641'121 (9%)

LL_RS7_G1_3	High	10'896'375	37'720'580	22.41%	48'616'955 (91%)
	Low	1'443'847	3'470'712	29.38%	4'914'559 (9%)

For all samples, ~88% of cytosines' methylation state was called with high confidence on the *halleri* side and ~91% had high confidence on the *lyrata* side. Low confidence cytosines show a higher methylation proportion compared to the high confidence cytosines.

Arabidopsis kamchatica generation 4 synthetic, stress 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	HOT_syn4v1_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 bisulfite 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_3 had the highest for both. 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	116'087'829	hal	23'538'891 (20.3%)	21'858'540 (92.3%)
		lyr	19'204'196 (16.5%)	17'712'818 (92.2%)
LL_RS7K_G4_2	73'493'997	hal	29'200'843 (39.7%)	27'022'593 (92.5%)

		lyr	25'273'175 (34.4%)	23'289'371 (92.2%)
LL_RS7K_G4_3	61'419'081	hal	26'001'701 (42.3%)	24'266'491 (93.3%)
		lyr	21'357'503 (34.8%)	19'784'065 (92.6%)

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	69.8%	59.6%	48.2%
	lyr	60.1%	48.8%	37.0%
LL_RS7K_G4_2	hal	79.5%	72.2%	61.1%
	lyr	70.7%	62.3%	50.0%
LL_RS7K_G4_3	hal	75.6%	67.6%	56.2%
	lyr	61.7%	49.9%	36.3%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_RS7K_G4_1	High	14'801'654	45'183'011	24.68%	59'984'665 (83%)
	Low	3'053'866	9'116'653	25.09%	12'170'519 (17%)
LL_RS7K_G4_2	High	15'667'666	47'496'966	24.80%	63'164'632 (88%)
	Low	2'427'418	6'563'134	27.00%	8'990'552 (12%)
LL_RS7K_G4_3	High	15'086'060	45'937'805	24.72%	61'023'865 (85%)
	Low	2'765'877	8'365'442	24.85%	11'131'319 (15%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)

LL_RS7K_G4_1	High	10'319'651	36'513'540	22.03%	46'833'191 (88%)
	Low	1'809'055	4'889'268	27.01%	6'698'323 (12%)
LL_RS7K_G4_2	High	10'785'502	37'883'807	22.16%	48'669'309 (91%)
	Low	1'434'158	3'428'047	29.50%	4'862'205 (9%)
LL_RS7K_G4_3	High	10'140'236	35'736'235	22.10%	45'876'471 (86%)
	Low	1'846'378	5'808'665	24.12%	7'655'043 (14%)

For all samples, ~85% of cytosines' methylation state was called with high confidence on the *halleri* side and ~90% 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), stress 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	HOT_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 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 bisulfite 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	65'488'677	hal	21'675'188 (33.1%)	21'056'750 (97.2%)
		lyr	18'698'934 (28.6%)	18'102'752 (96.8%)
LL_TKS_G1_2	63'982'146	hal	20'765'997 (32.5%)	20'205'350 (97.3%)
		lyr	18'387'646 (28.7%)	17'840'905 (97.0%)
LL_TKS_G1_3	63'923'640	hal	21'085'681 (33.0%)	20'479'971 (97.1%)
		lyr	18'148'375	17'574'113

			(28.4%)	(96.8%)

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	64.6%	58.1%	48.0%
	lyr	53.9%	45.5%	34.6%
LL_TKS_G1_2	hal	64.0%	57.3%	47.2%
	lyr	53.5%	45.0%	34.0%
LL_TKS_G1_3	hal	63.4%	55.8%	44.8%
	lyr	52.6%	43.3%	31.9%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_TKS_G1_1	High	12'504'782	41'788'859	23.03%	54'293'641 (75%)
	Low	3'642'256	14'219'287	20.39%	17'861'543 (25%)
LL_TKS_G1_2	High	12'675'113	41'412'390	23.43%	54'087'503 (75%)
	Low	3'765'967	14'301'714	20.84%	18'067'681 (25%)
LL_TKS_G1_3	High	12'553'600	41'390'727	23.27%	53'944'327 (75%)
	Low	3'726'801	14'484'056	20.46%	18'210'857 (25%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
LL_TKS_G1_1	High	7'729'927	30'667'304	20.13%	38'397'231 (72%)
	Low	2'851'370	12'282'913	18.84%	15'134'283 (28%)
LL_TKS_G1_2	High	7'858'631	30'491'355	20.49%	38'349'986 (72%)

	Low	2'928'281	12'253'247	19.29%	15'181'528 (28%)
LL_TKS_G1_3	High	7'767'388	30'316'073	20.40%	38'083'461 (71%)
	Low	2'858'828	12'589'225	18.51%	15'448'053 (29%)

For all samples, ~73% 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), stress 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	HOT_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 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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after read sorting was similar between samples, ranging between 28-33%. 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	61'314'455	hal	20'195'034 (32.9%)	19'613'044 (97.1%)
		lyr	17'700'780 (28.9%)	17'105'743 (96.6%)
LL_TKS_G5_2	92'014'805	hal	31'558'736 (34.3%)	30'699'977 (97.3%)
		lyr	28'217'031 (30.7%)	27'310'101 (96.8%)
LL_TKS_G5_3	77'008'782	hal	24'928'590 (32.4%)	24'137'298 (96.8%)

		lyr	21'791'774 (28.3%)	20'978'045 (96.3%)

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_G5_1	hal	63.1%	55.2%	43.9%
	lyr	52.1%	42.5%	30.9%
LL_TKS_G5_2	hal	67.1%	62.8%	57.5%
	lyr	58.3%	52.2%	45.2%
LL_TKS_G5_3	hal	65.4%	59.7%	50.9%
	lyr	55.2%	47.6%	37.6%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_TKS_G5_1	High	12'132'265	41'435'379	22.65%	53'567'644 (74%)
	Low	3'803'578	14'783'962	20.46%	18'587'540 (26%)
LL_TKS_G5_2	High	13'069'747	43'082'067	23.28%	56'151'814 (78%)
	Low	3'490'409	12'512'961	21.81%	16'003'370 (22%)
LL_TKS_G5_3	High	12'398'543	42'034'201	22.78%	54'432'744 (75%)
	Low	3'583'744	14'138'696	20.22%	17'722'440 (25%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
LL_TKS_G5_1	High	7'480'601	30'269'250	19.82%	37'749'851 (70%)
	Low	2'940'946	12'840'717	18.64%	15'781'663 (30%)
LL_TKS_G5_2	High	8'508'087	32'373'628	20.81%	40'881'715 (76%)
	Low	2'684'543	9'965'256	21.22%	12'649'799 (24%)

LL_TKS_G5_3	High	7'674'921	30'823'970	19.94%	38'498'891 (72%)
	Low	2'782'066	12'250'557	18.51%	15'032'623 (28%)

For all samples, ~74% 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), stress 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	HOT_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
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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after read sorting was similar between samples. Deduplication rates were around 6%, 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	92'607'805	hal	30'566'880 (33.0%)	28'760'149 (94.1%)
		lyr	28'748'016 (31.0%)	26'912'357 (93.6%)
LL_ALK_G1_2	84'880'448	hal	29'241'368 (34.5%)	27'436'681 (93.8%)
		lyr	27'363'037 (32.2%)	25'562'083 (93.4%)
LL_ALK_G1_3	75'710'886	hal	23'379'684 (30.9%)	21'837'094 (93.4%)

		lyr	22'202'653 (29.3%)	20'664'506 (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	66.8%	62.4%	56.4%
	lyr	61.8%	56.1%	48.8%
LL_ALK_G1_2	hal	66.3%	61.7%	55.3%
	lyr	61.2%	55.3%	47.6%
LL_ALK_G1_3	hal	64.9%	59.7%	51.7%
	lyr	59.4%	52.6%	43.2%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_ALK_G1_1	High	13'413'806	41'991'643	24.21%	55'405'449 (77%)
	Low	3'496'369	13'253'366	20.87%	16'749'735 (23%)
LL_ALK_G1_2	High	13'354'101	41'682'968	24.26%	55'037'069 (76%)
	Low	3'609'613	13'508'502	21.09%	17'118'115 (24%)
LL_ALK_G1_3	High	13'294'636	41'088'167	24.45%	54'382'803 (75%)
	Low	3'773'438	13'998'943	21.23%	17'772'381 (25%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
LL_ALK_G1_1	High	9'000'107	32'158'586	21.87%	41'158'693 (77%)
	Low	2'549'173	9'823'648	20.60%	12'372'821 (23%)
LL_ALK_G1_2	High	8'950'072	31'856'451	21.93%	40'806'523 (76%)
	Low	2'620'257	10'104'734	20.59%	12'724'991 (24%)

LL_ALK_G1_3	High	8'847'131	31'345'743	22.01%	40'192'874 (75%)
	Low	2'716'236	10'622'404	20.36%	13'338'640 (25%)

For all samples, ~76% 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), stress 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	HOT_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
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 bisulfite conversion and library preparation.

Alignment and deduplication

The mapping rate after read sorting was similar between samples. 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	101'404'674	hal	34'939'757 (34.5%)	32'392'865 (92.7%)
		lyr	33'831'750 (33.4%)	31'141'162 (92.1%)
LL_ALK_G4_2	83'724'490	hal	26'673'517 (31.9%)	24'669'377 (92.5%)
		lyr	25'230'328 (30.1%)	23'204'118 (92.0%)
LL_ALK_G4_3	64'454'293	hal	21'574'308 (33.5%)	19'964'514 (92.5%)

		lyr	20'286'838 (31.5%)	18'688'439 (92.1%)

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 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	67.0%	63.1%	58.6%
	lyr	62.3%	57.3%	51.5%
LL_ALK_G4_2	hal	65.2%	60.4%	53.5%
	lyr	59.4%	52.6%	43.9%
LL_ALK_G4_3	hal	62.8%	55.3%	44.2%
	lyr	56.8%	47.9%	36.0%

Confidence of methylation calls

halleri	Confidence	Methylated	Not methylated	%	Total (%)
LL_ALK_G4_1	High	13'634'864	41'501'748	24.73%	55'136'612 (76%)
	Low	3'624'517	13'394'055	21.30%	17'018'572 (24%)
LL_ALK_G4_2	High	12'977'571	41'279'976	23.92%	54'257'547 (75%)
	Low	3'675'968	14'221'669	20.54%	17'897'637 (25%)
LL_ALK_G4_3	High	12'582'537	40'937'220	23.51%	53'519'757 (74%)
	Low	3'860'248	14'775'179	20.71%	18'635'427 (26%)

lyrata	Confidence	Methylated	Not methylated	%	Total (%)
LL_ALK_G4_1	High	9'153'139	31'838'081	22.33%	40'991'220 (81%)
	Low	2'629'463	9'910'831	20.97%	12'540'294 (19%)
LL_ALK_G4_2	High	8'678'088	31'441'018	21.63%	40'119'106 (80%)

	Low	2'686'004	10'726'404	20.03%	13'412'408 (20%)
LL_ALK_G4_3	High	8'408'370	31'107'323	21.28%	39'515'693 (79%)
	Low	2'758'560	11'257'261	19.68%	14'015'821 (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.