DNBSEQ-G400 WGS variant analysis evaluation and comparison with NA19240 sample-matched Illumina data from SRA

Stéphane Plaisance [VIB - Nucleomics Core, nucleomics@vib.be]

February 21st, 2020 - version 1.0

Contents

last edits: Fri Mar 06, 2020

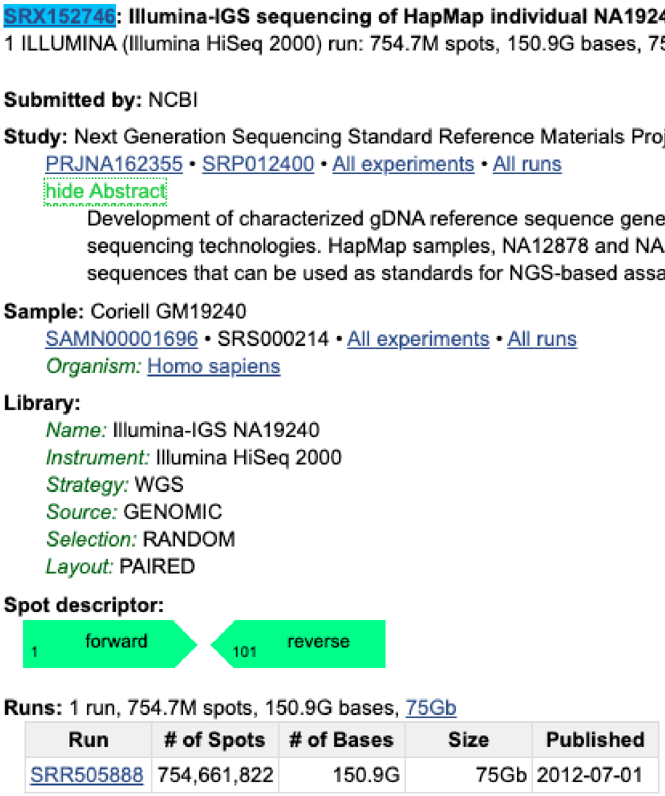
Introduction

This work compares results obtained from a run on our new DNBSeq-G400 MGI sequencer and Illumina paired-end read data for the same genomes obtained from the SRA repository. The amount of data in both sequencing runs is comparable and corresponds to roughly 30x human genome coverage.

The different stages of the GATK Best Practice workflow applied here show very similar results for both data-sets and confirm that the MGI platform produces paired reads with equal quality to the classical Illumina platforms.

This work aims at comparing GATK4 variant calls obtained from the training run (lane L04 out of 4 lanes ran on the machine during the training week) and public data for the same 1000 genome sample YRI daughter **NA19240 Coriell link**

The compared sample is part of "PRJNA200694" : A public-private-academic consortium, Genome-in-a-Bottle (GIAB), hosted by NIST to develop reference materials and standards for clinical sequencing **PRJNA162355 link** (which also comprises 3 other genomes).



SRX152746: Illumina-IGS sequencing of HapMap individual NA19240
1 ILLUMINA (Illumina HiSeq 2000) run: 754.7M spots, 150.9G bases, 75Gb downloads

Submitted by: NCBI

Study: Next Generation Sequencing Standard Reference Materials Project
PRJNA162355 • SRP012400 • All experiments • All runs
hide Abstract
Development of characterized gDNA reference sequence generated by various Next Generation sequencing technologies. HapMap samples, NA12878 and NA19240 will be used to generate reference sequences that can be used as standards for NGS-based assays

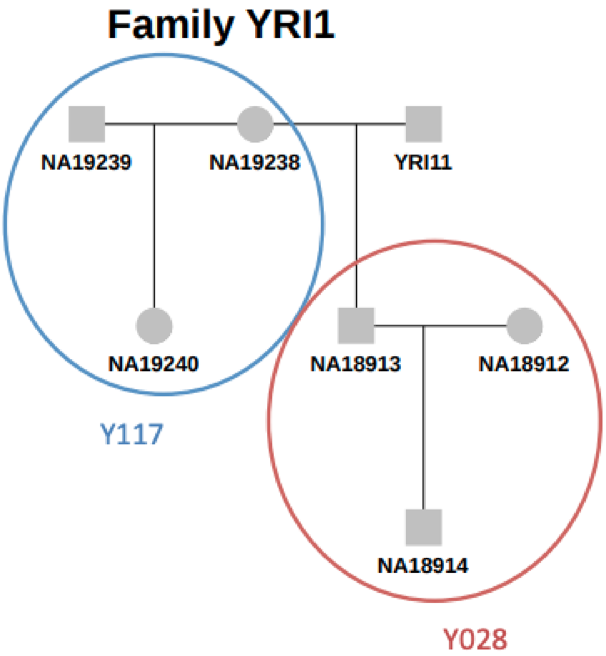
Sample: Coriell GM19240
SAMN00001696 • SRS000214 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: Illumina-IGS NA19240
Instrument: Illumina HiSeq 2000
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 754.7M spots, 150.9G bases, 75Gb

Run	# of Spots	# of Bases	Size	Published
SRR505888	754,661,822	150.9G	75Gb	2012-07-01



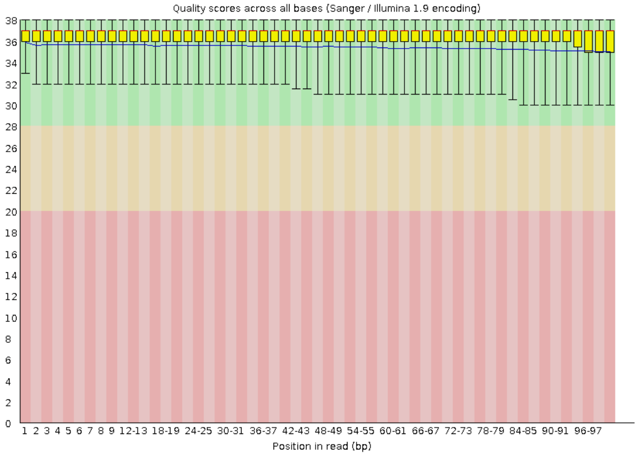
Results

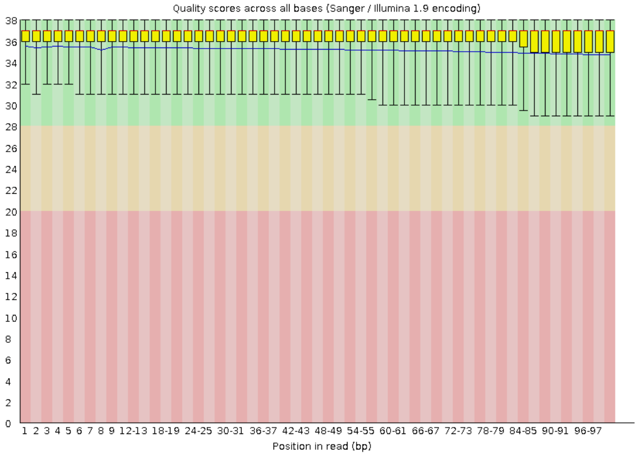
Raw reads were submitted to FastQC in order to identify intrinsic biases including left-over adapters or overabundant kmers. Base-calling quality was assessed across the whole read length as done routinely.

Read QC using FastQC

FastQC Results for the local MGI reads:

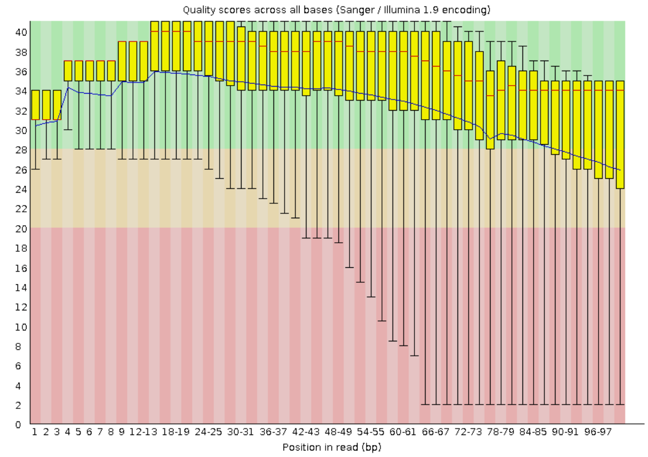
The overall quality seems biased towards 38 and stays very high across the whole cycle range. This is probably a specificity of the base-calling method used by MGI which delivers very constant and high values for all bases.

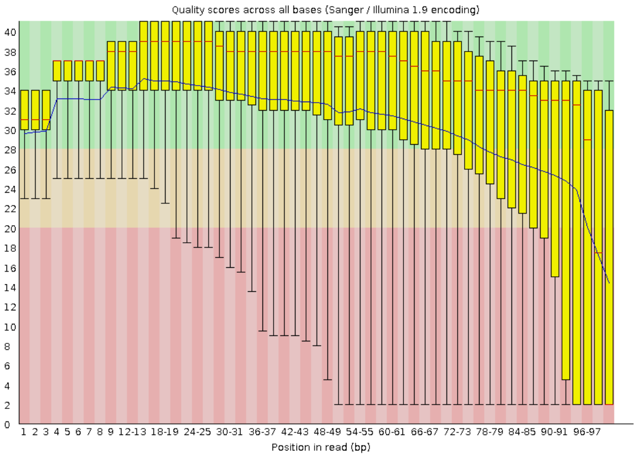




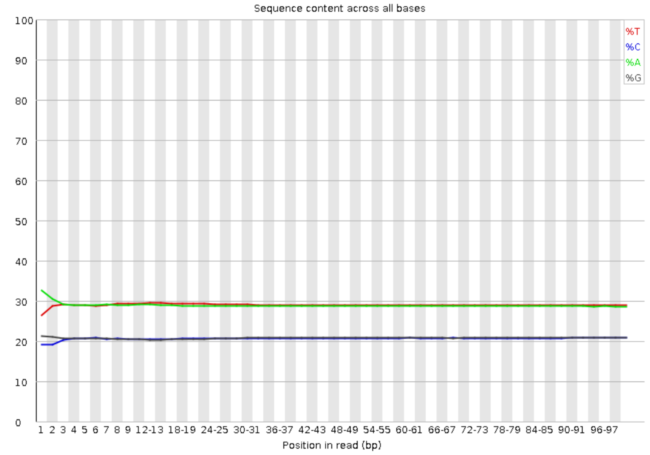
FastQC Results for the SRA Illumina reads:

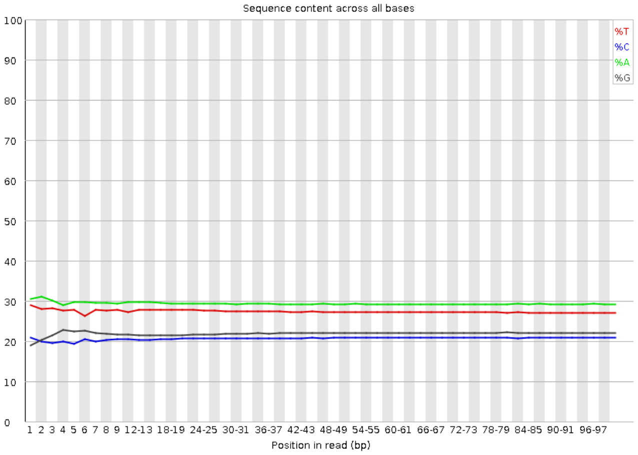
The base-call quality is generally lower with these reads and drops significantly at the right (higher cycle numbers) of the plots



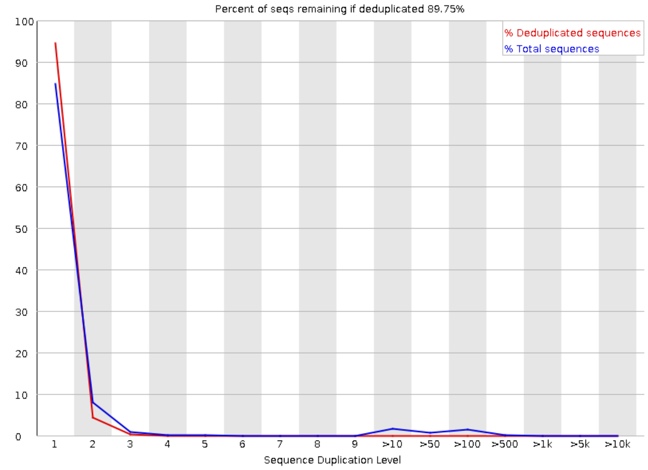


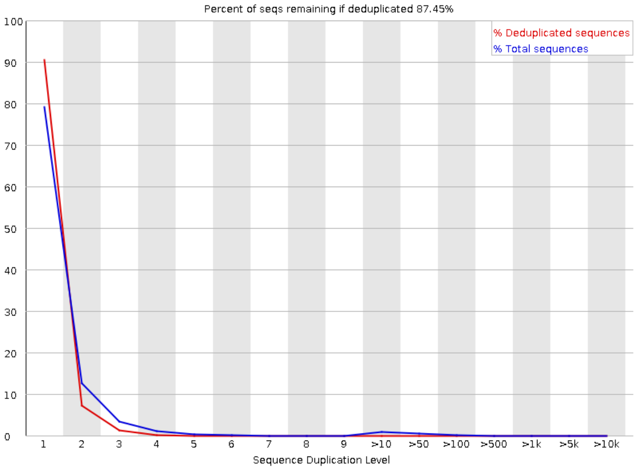
Base composition is the same fore both data-sets (here only reads_1). The left plot is for the MGI data and the right plot for the SRA data.



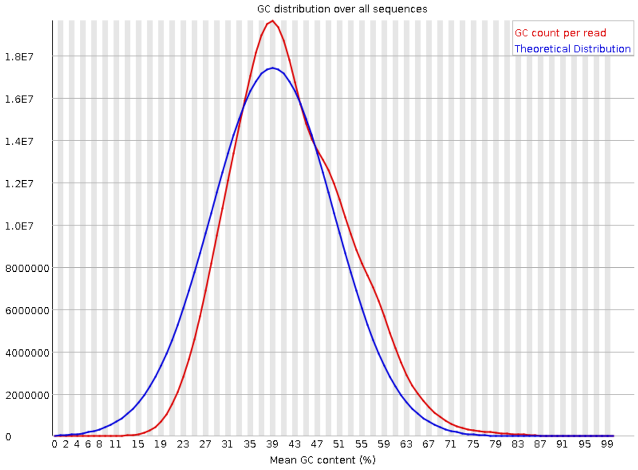


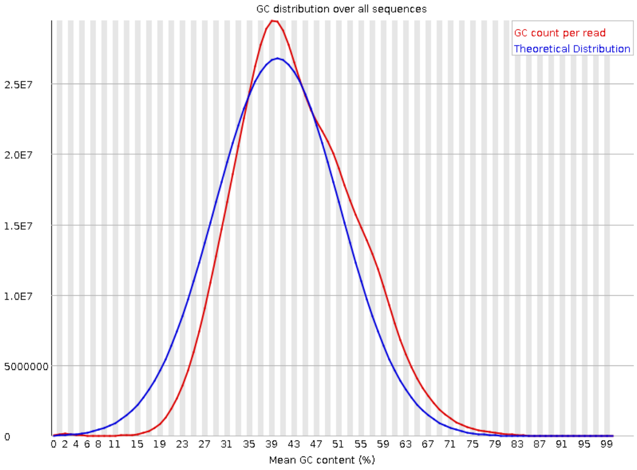
The sequence duplication level is also very comparable (here only reads_1). The left plot is for the MGI data and the right plot for the SRA data.





The per sequence GC content is also very comparable (here only reads_1). The left plot is for the MGI data and the right plot for the SRA data.





There were no over-expressed kmers detected in either data-set.

Mapping reads to GRCh38 with BWA mem

A number of files were obtained from the **Broad Bundle FTP repository** to be used in with GATK. For read mapping, a BWA index was build using the reference **Homo_sapiens_assembly38** from the bundle.

BWA mapping resulted in a very high proportion of paired-reads mapped to the reference for both data-sets and a slightly better score for the MGI data (97.73% for MGI reads vs 93.16% for SRA reads)

Samtools flagstats for the MGI data:

```
959600075 + 0 in total (QC-passed reads + QC-failed reads)
2296547 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
958489133 + 0 mapped (99.88% : N/A)
957303528 + 0 paired in sequencing
478651764 + 0 read1
478651764 + 0 read2
935619392 + 0 properly paired (97.73% : N/A)
955296958 + 0 with itself and mate mapped
895628 + 0 singletons (0.09% : N/A)
15434370 + 0 with mate mapped to a different chr
11850427 + 0 with mate mapped to a different chr (mapQ>=5)
```

The raw coverage depth value obtained from **479'800'038** 100bp read-pairs is **30.0 x**

Samtools flagstats for the SRA data:

```
1513475409 + 0 in total (QC-passed reads + QC-failed reads)
4151765 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
1472588400 + 0 mapped (97.30% : N/A)
1509323644 + 0 paired in sequencing
754661822 + 0 read1
754661822 + 0 read2
1406113578 + 0 properly paired (93.16% : N/A)
1454423486 + 0 with itself and mate mapped
14013149 + 0 singletons (0.93% : N/A)
22424390 + 0 with mate mapped to a different chr
11120027 + 0 with mate mapped to a different chr (mapQ>=5)
```

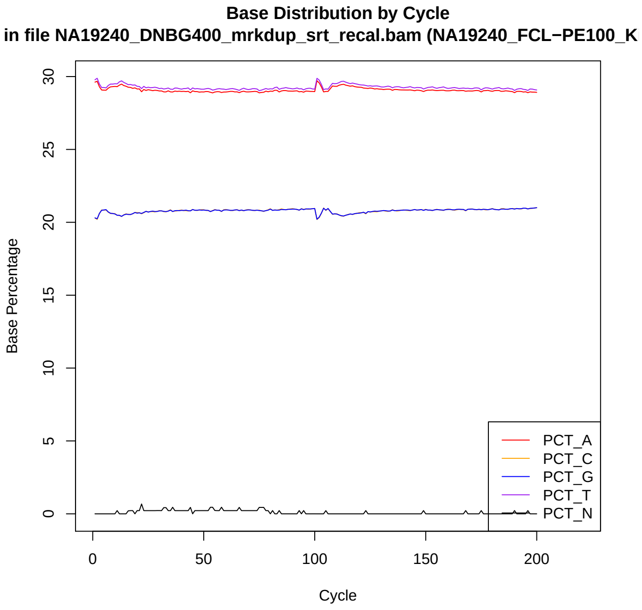
The raw coverage depth value obtained from **756'737'704** 100bp read-pairs is **47.3 x** which gives a significant advantage to the SRA data-set.

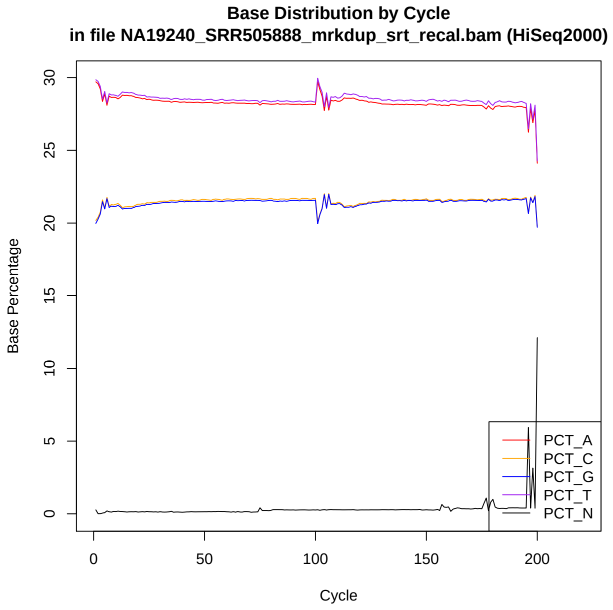
GATK4 pre-processing and BQSR

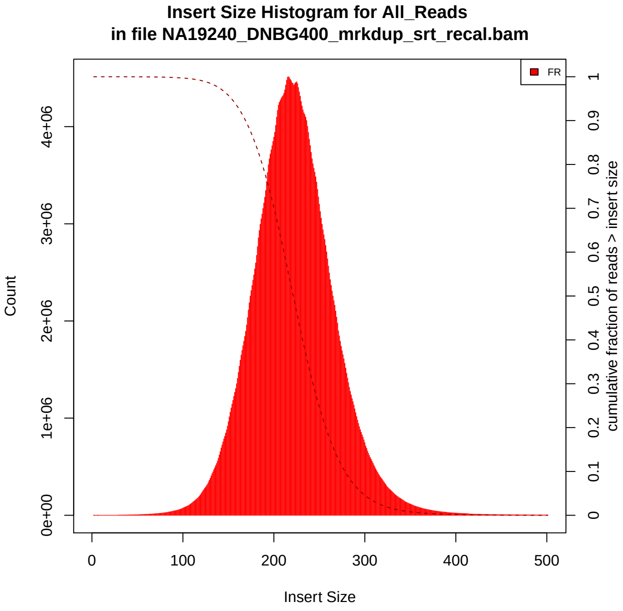
Both raw mapping sets were pre-processed as described in the GATK Best practices in order to **mark read duplicates**, **fix erroneous SAM flags**, and **perform base-call quality score re-calibration (BQSR)** resulting in a final BAM file for each platform.

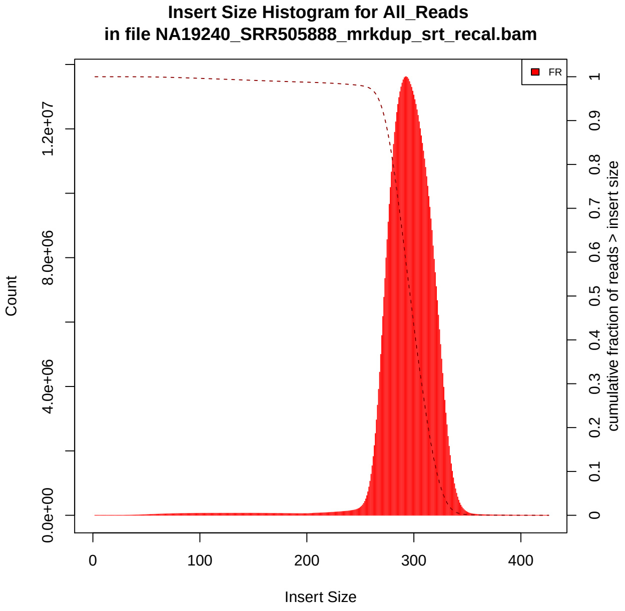
The final BAM data was assessed using two QC tools: Picard CollectMultipleMetrics and Qualimap

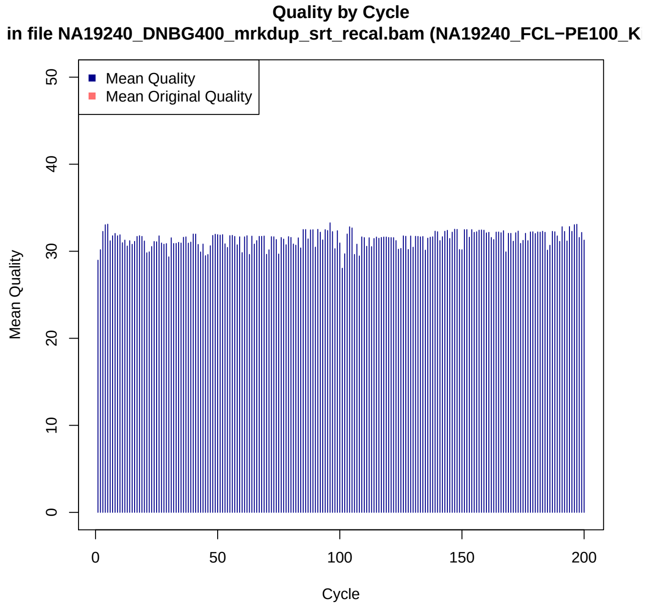
Picard CollectMultipleMetrics QC results

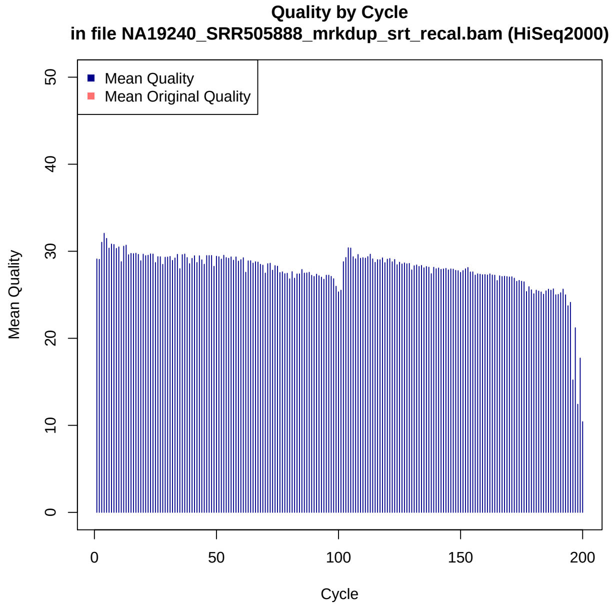


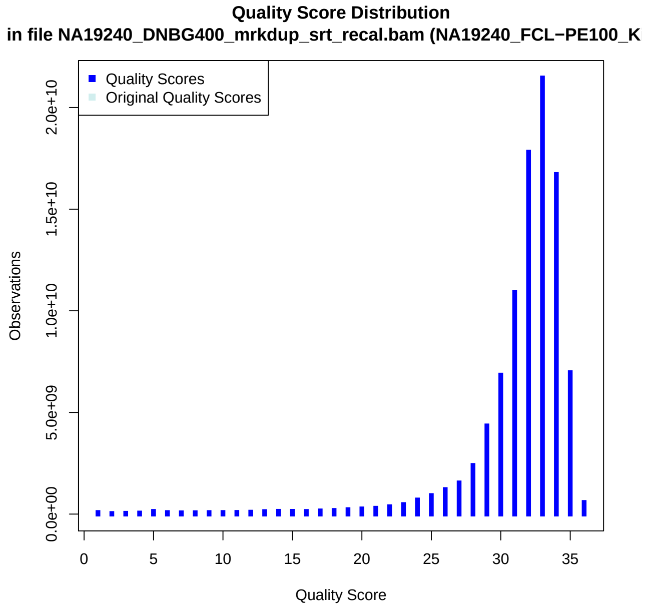


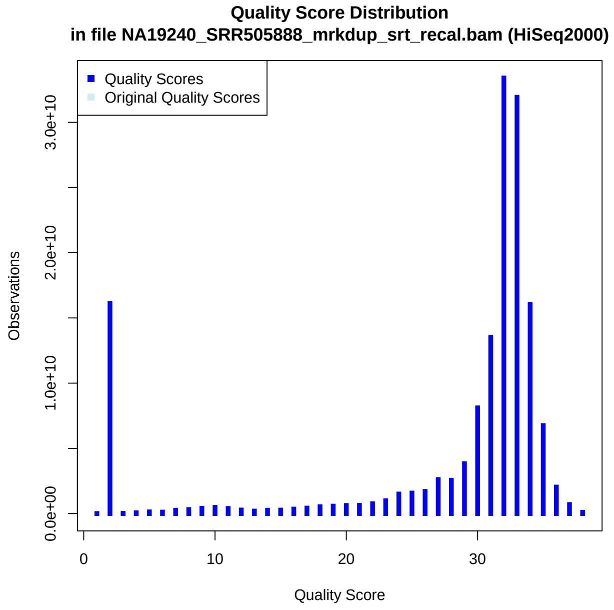






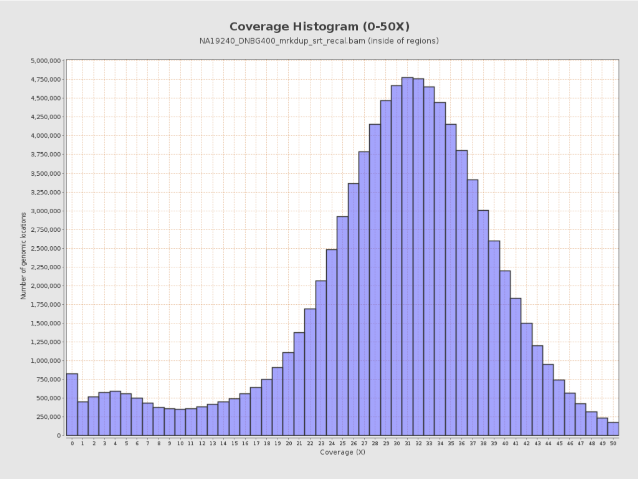






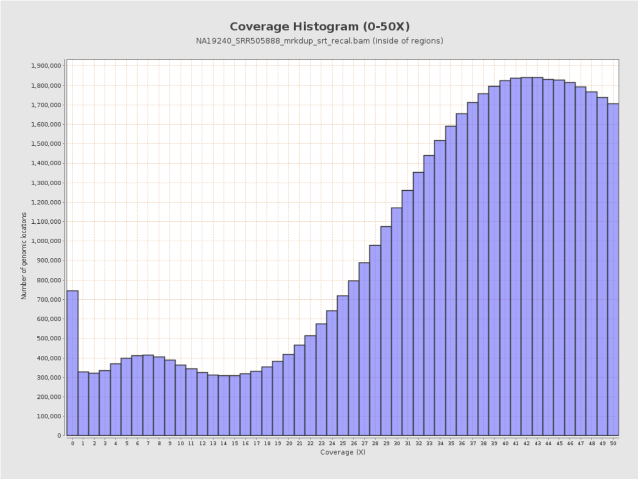
Qualimap QC results

All Qualimap results are reported side by side below. The left plot of each pair is from the local MGI sequencing set.



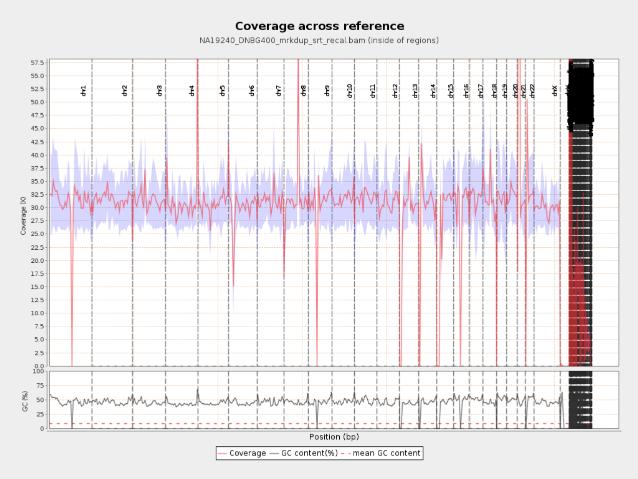
Coverage Histogram (0-50X)

NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)



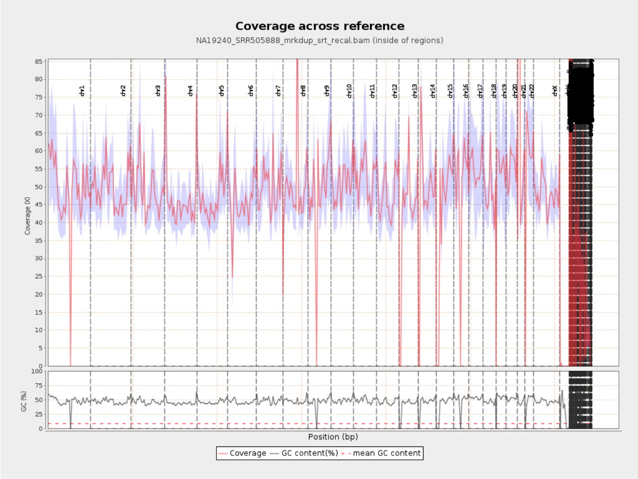
Coverage Histogram (0-50X)

NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)



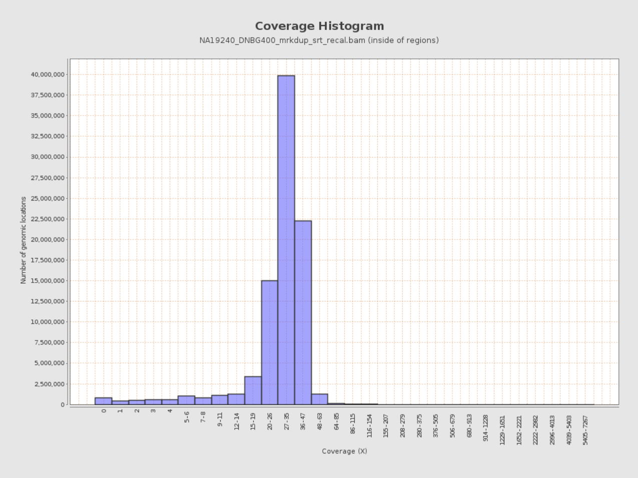
Coverage across reference

NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)



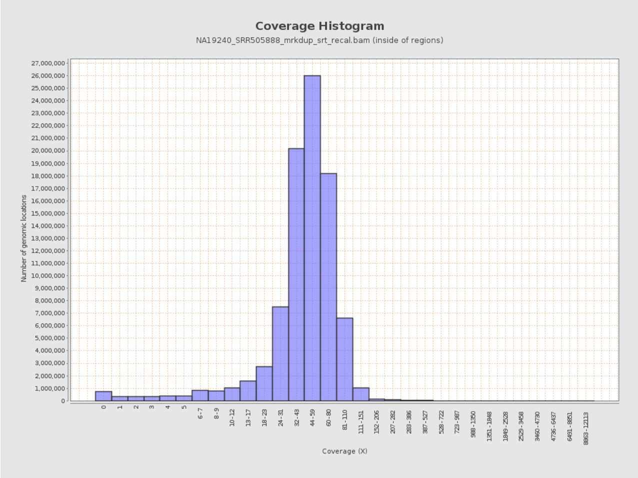
Coverage across reference

NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)



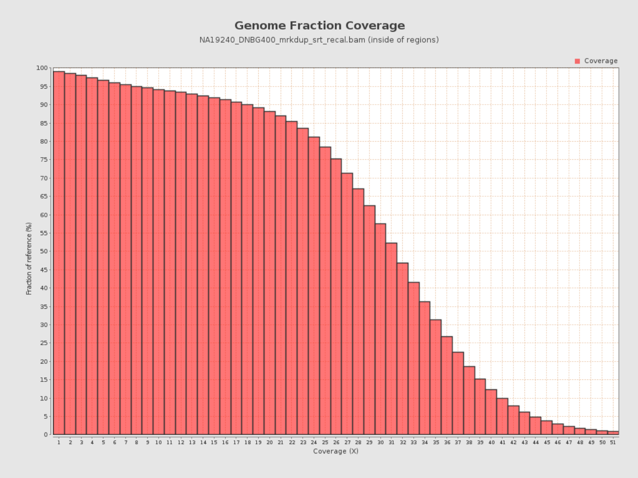
Coverage Histogram

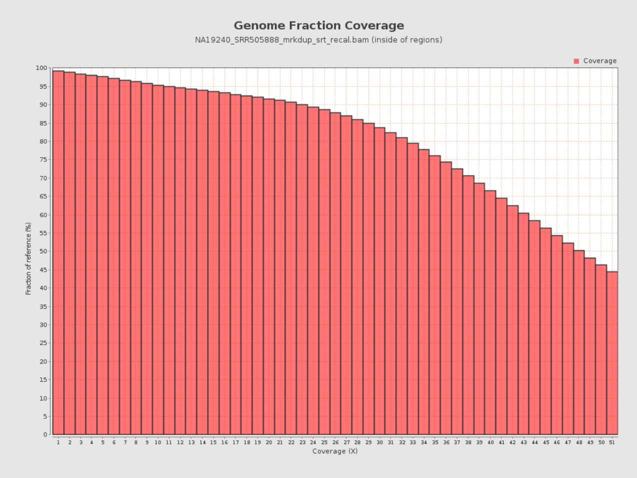
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)

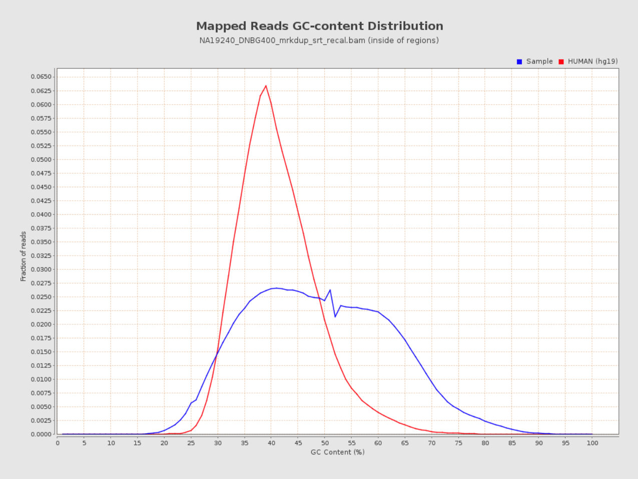


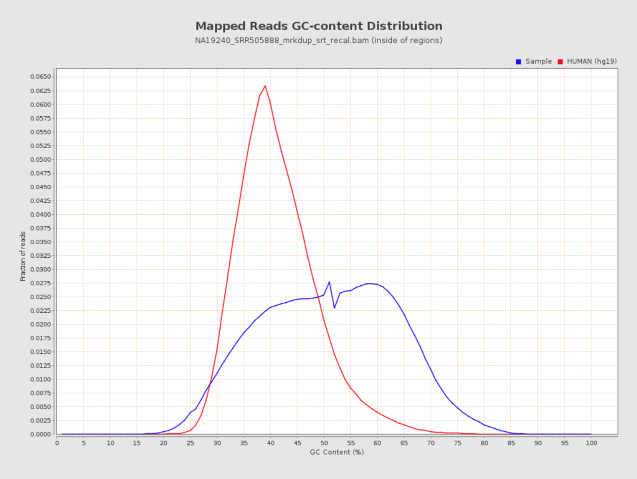
Coverage Histogram

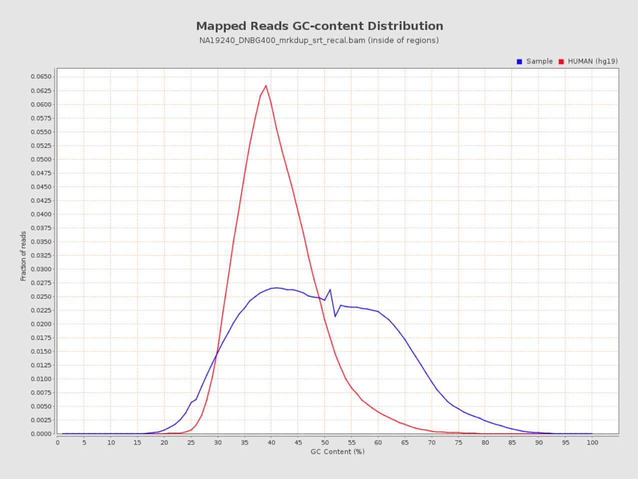
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)

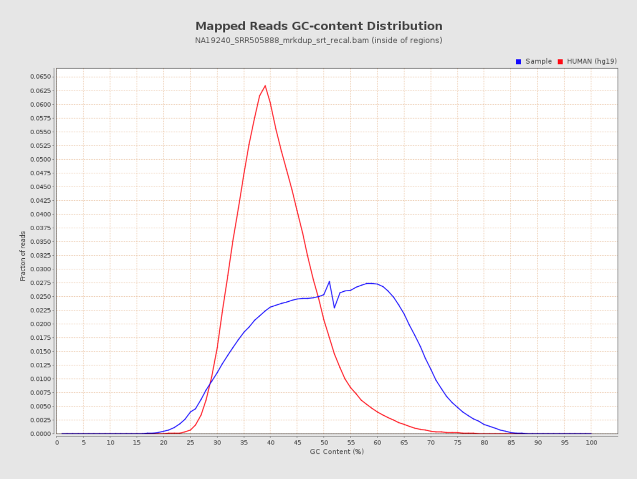

Genome Fraction Coverage
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
Coverage
Fraction of reference (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Coverage (X)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51

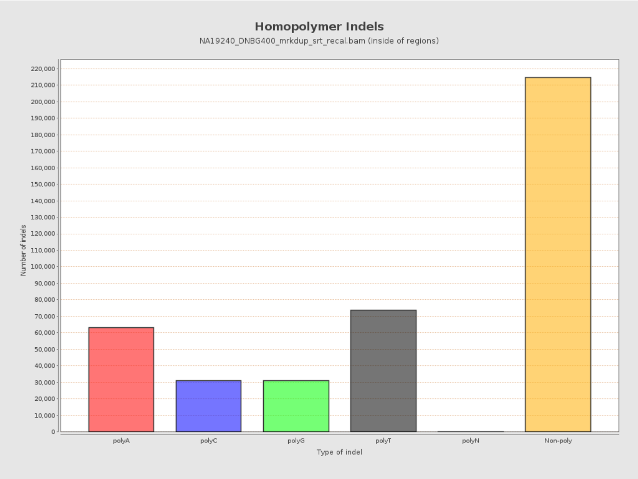

Genome Fraction Coverage
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
Coverage
Fraction of reference (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Coverage (X)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51


Mapped Reads GC-content Distribution
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
Sample
HUMAN (hg19)
Fraction of reads
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200 0.0225 0.0250 0.0275 0.0300 0.0325 0.0350 0.0375 0.0400 0.0425 0.0450 0.0475 0.0500 0.0525 0.0550 0.0575 0.0600 0.0625 0.0650
GC Content (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100


Mapped Reads GC-content Distribution
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
Sample
HUMAN (hg19)
Fraction of reads
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200 0.0225 0.0250 0.0275 0.0300 0.0325 0.0350 0.0375 0.0400 0.0425 0.0450 0.0475 0.0500 0.0525 0.0550 0.0575 0.0600 0.0625 0.0650
GC Content (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

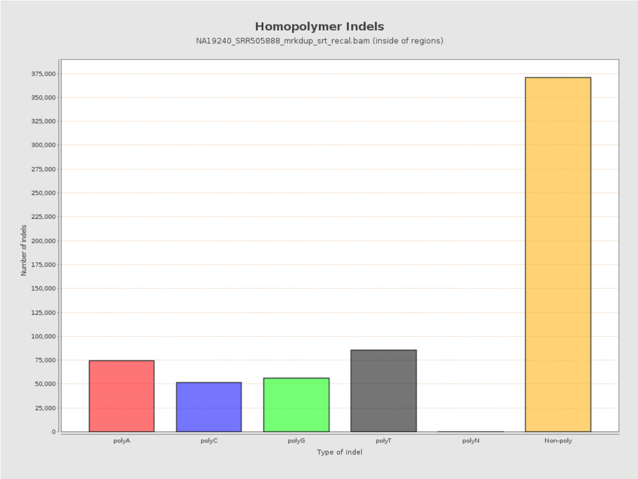

Mapped Reads GC-content Distribution
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
Sample
HUMAN (hg19)
Fraction of reads
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200 0.0225 0.0250 0.0275 0.0300 0.0325 0.0350 0.0375 0.0400 0.0425 0.0450 0.0475 0.0500 0.0525 0.0550 0.0575 0.0600 0.0625 0.0650
GC Content (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100


Mapped Reads GC-content Distribution
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
Sample
HUMAN (hg19)
Fraction of reads
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200 0.0225 0.0250 0.0275 0.0300 0.0325 0.0350 0.0375 0.0400 0.0425 0.0450 0.0475 0.0500 0.0525 0.0550 0.0575 0.0600 0.0625 0.0650
GC Content (%)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100



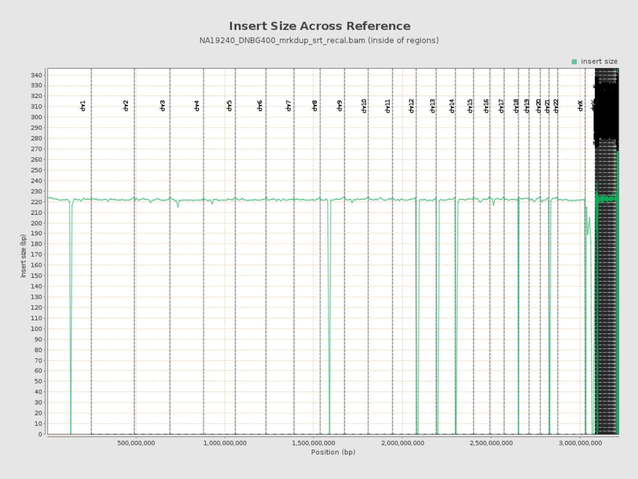
Homopolymer Indels

NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)



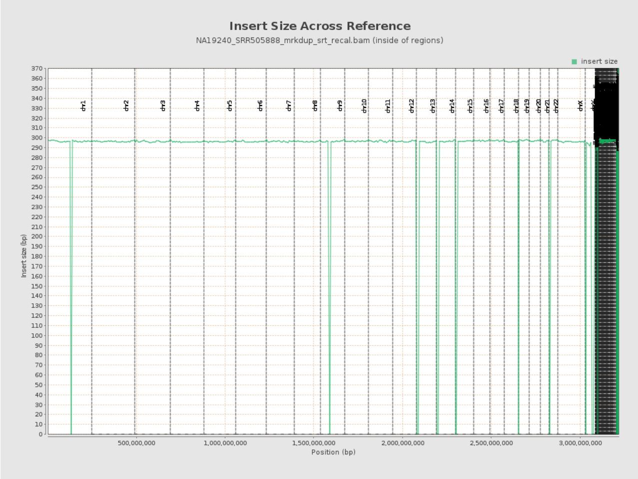
Homopolymer Indels

NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)



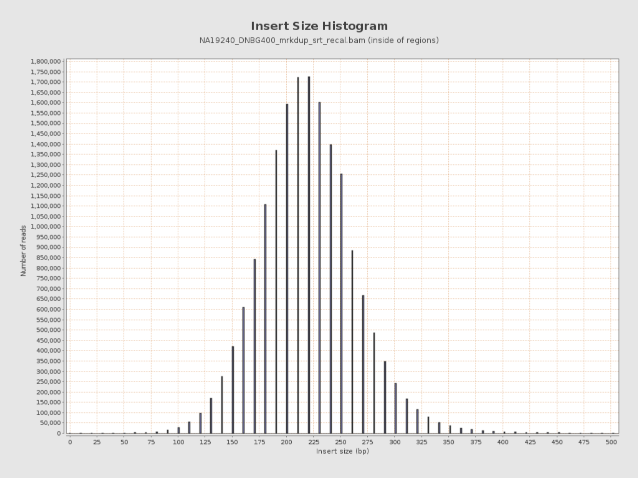
Insert Size Across Reference

NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)



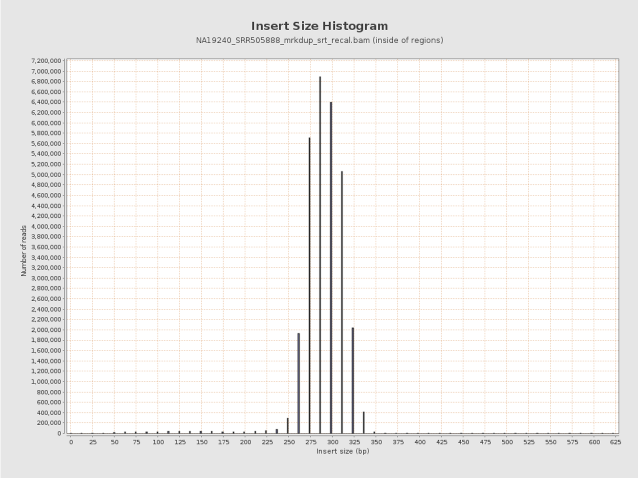
Insert Size Across Reference

NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)



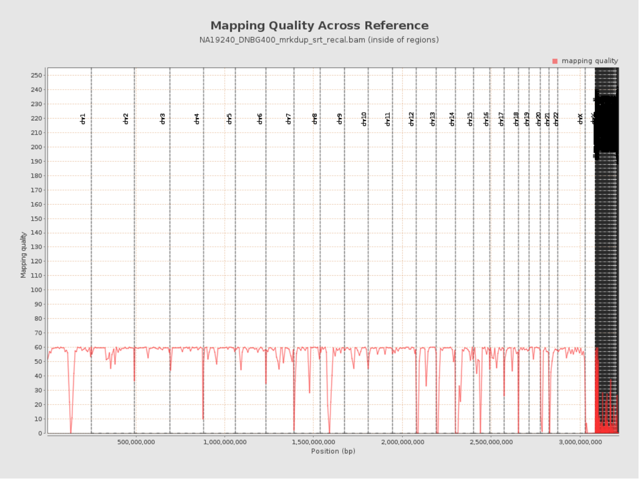
Insert Size Histogram

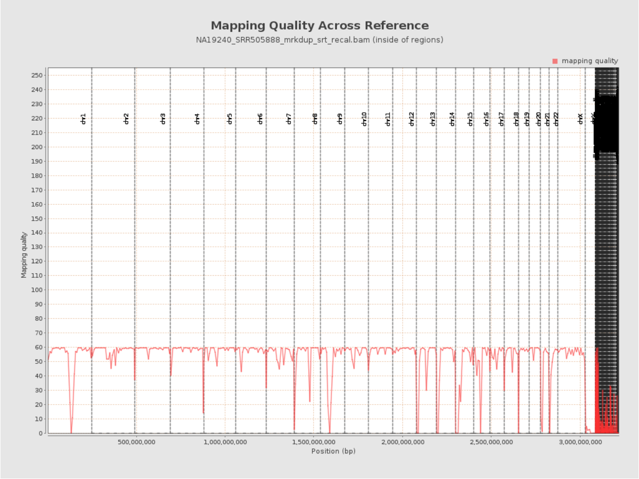
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)

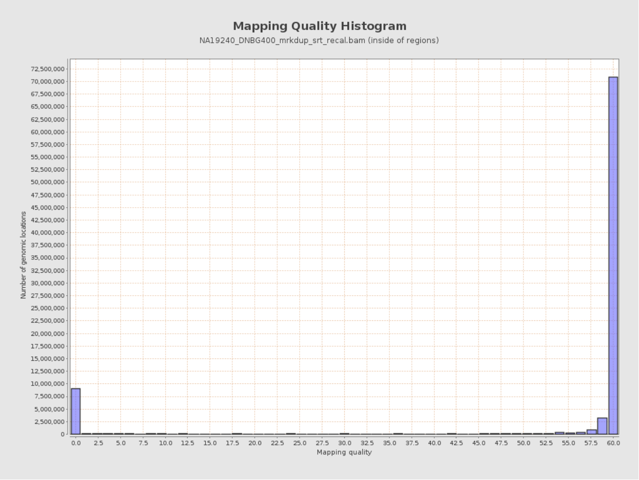


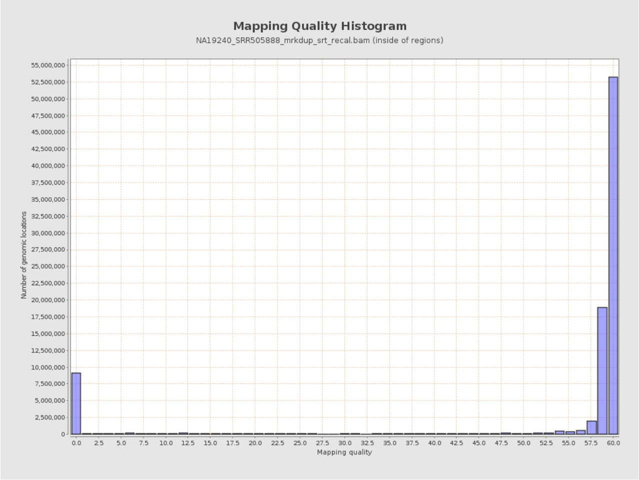
Insert Size Histogram

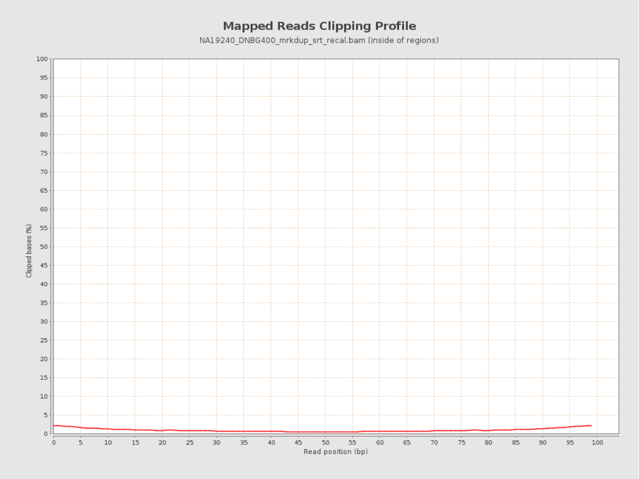
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)

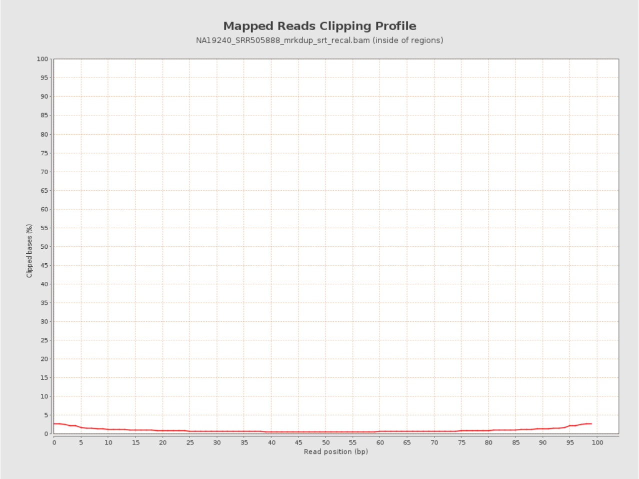

Mapping Quality Across Reference
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
mapping quality
Mapping quality
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
500,000,000
1,000,000,000
1,500,000,000
2,000,000,000
2,500,000,000
3,000,000,000
Position (bp)

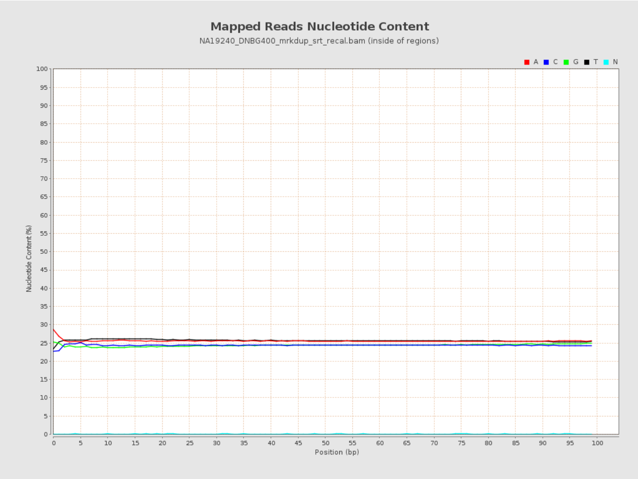

Mapping Quality Across Reference
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
mapping quality
Mapping quality
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
500,000,000
1,000,000,000
1,500,000,000
2,000,000,000
2,500,000,000
3,000,000,000
Position (bp)

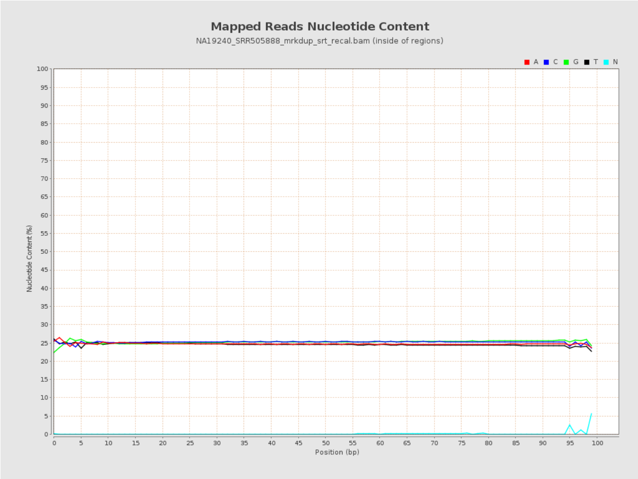

Mapping Quality Histogram
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
Number of genomic locations
0
2,500,000
5,000,000
7,500,000
10,000,000
12,500,000
15,000,000
17,500,000
20,000,000
22,500,000
25,000,000
27,500,000
30,000,000
32,500,000
35,000,000
37,500,000
40,000,000
42,500,000
45,000,000
47,500,000
50,000,000
52,500,000
55,000,000
57,500,000
60,000,000
62,500,000
65,000,000
67,500,000
70,000,000
72,500,000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
Mapping quality

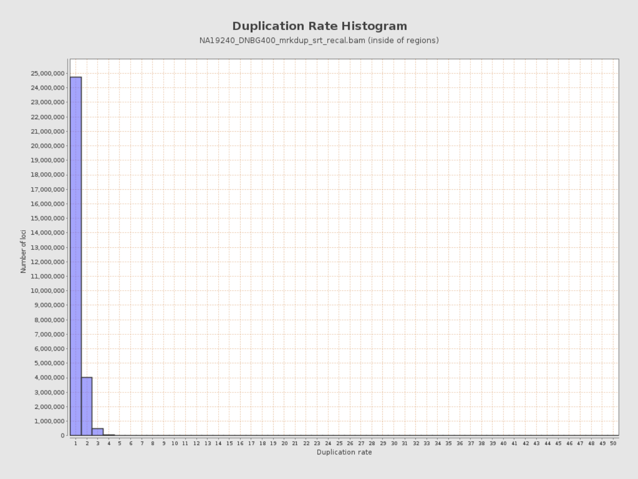

Mapping Quality Histogram
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
Number of genomic locations
0
2,500,000
5,000,000
7,500,000
10,000,000
12,500,000
15,000,000
17,500,000
20,000,000
22,500,000
25,000,000
27,500,000
30,000,000
32,500,000
35,000,000
37,500,000
40,000,000
42,500,000
45,000,000
47,500,000
50,000,000
52,500,000
55,000,000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0
42.5
45.0
47.5
50.0
52.5
55.0
57.5
60.0
Mapping quality

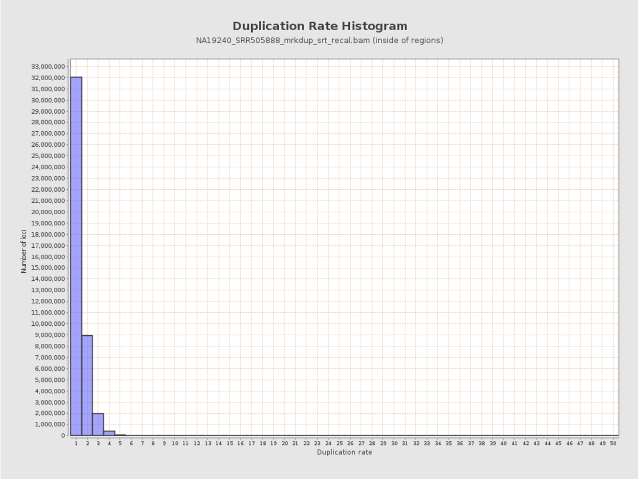

Mapped Reads Clipping Profile
NA19240_DNBG400_mrkdup_srt_recal.bam (inside of regions)
Clipped bases (%)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Read position (bp)


Mapped Reads Clipping Profile
NA19240_SRR505888_mrkdup_srt_recal.bam (inside of regions)
Clipped bases (%)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Read position (bp)





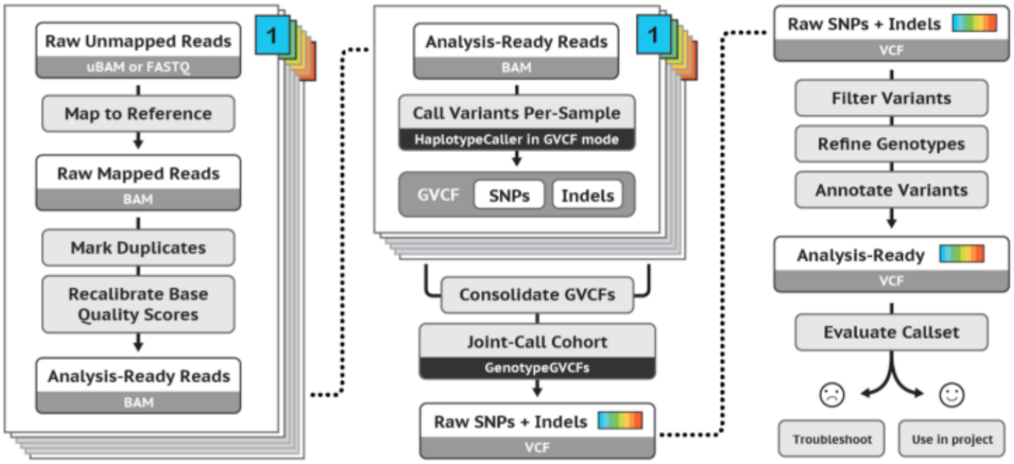




REM: The significant numerical advantage of the SRA data is confirmed by a number of plots. It would have been better to subset the SRA data to retain identical number of reads as in the MGI data-set but the final results will show that this was not necessary to come to conclusions.

GATK4 HaplotypeCaller germline SNP and Indel calling from pre-processed mappings

We next followed the Best Practice Workflow 'Germline short variant discovery SNPs Indels' as described on the **GATK pages**

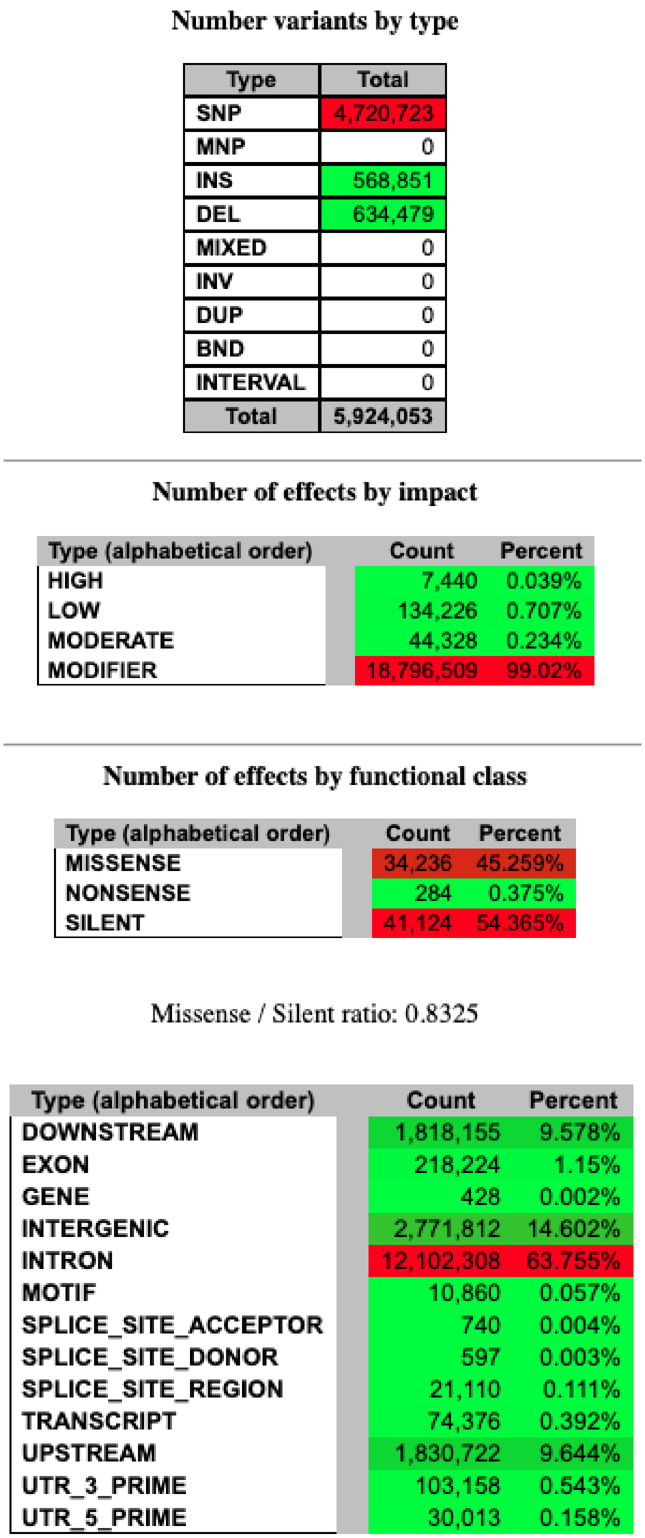


Each data-set was called using the HaplotypeCaller (v4.1.4.1 in GVCF mode) and the resulting files further processed as in the practice to lead to re-calibrated variant files. The VCF files were annotated using SNPEff to produce featured metrics shown below.

REM: The calling process is ran in parallel using sparks while some of the subsequent steps run in single core mode and are quite lengthy. This makes the full calling process took for each genome more than two days on 88 cores.

SNPEff annotation of the variant files

Key SNPEff results are reported side by side below. The left plot of each pair is from the local MGI sequencing set.



Number variants by type

Type	Total
SNP	4,720,723
MNP	0
INS	568,851
DEL	634,479
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	5,924,053

Number of effects by impact

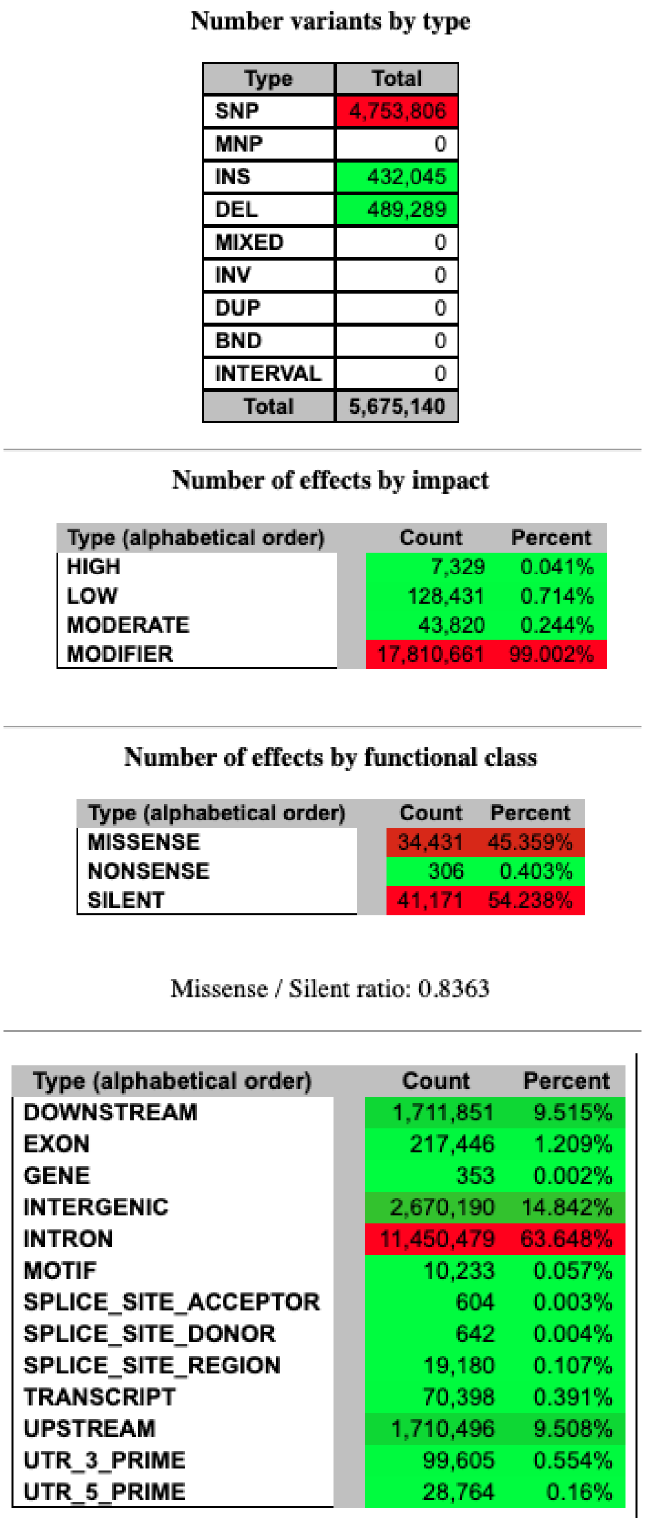
Type (alphabetical order)	Count	Percent
HIGH	7,440	0.039%
LOW	134,226	0.707%
MODERATE	44,328	0.234%
MODIFIER	18,796,509	99.02%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	34,236	45.259%
NONSENSE	284	0.375%
SILENT	41,124	54.365%

Missense / Silent ratio: 0.8325

Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,818,155	9.578%
EXON	218,224	1.15%
GENE	428	0.002%
INTERGENIC	2,771,812	14.602%
INTRON	12,102,308	63.755%
MOTIF	10,860	0.057%
SPLICE_SITE_ACCEPTOR	740	0.004%
SPLICE_SITE_DONOR	597	0.003%
SPLICE_SITE_REGION	21,110	0.111%
TRANSCRIPT	74,376	0.392%
UPSTREAM	1,830,722	9.644%
UTR_3_PRIME	103,158	0.543%
UTR_5_PRIME	30,013	0.158%



Number variants by type

Type	Total
SNP	4,753,806
MNP	0
INS	432,045
DEL	489,289
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	5,675,140

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	7,329	0.041%
LOW	128,431	0.714%
MODERATE	43,820	0.244%
MODIFIER	17,810,661	99.002%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	34,431	45.359%
NONSENSE	306	0.403%
SILENT	41,171	54.238%

Missense / Silent ratio: 0.8363

Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,711,851	9.515%
EXON	217,446	1.209%
GENE	353	0.002%
INTERGENIC	2,670,190	14.842%
INTRON	11,450,479	63.648%
MOTIF	10,233	0.057%
SPLICE_SITE_ACCEPTOR	604	0.003%
SPLICE_SITE_DONOR	642	0.004%
SPLICE_SITE_REGION	19,180	0.107%
TRANSCRIPT	70,398	0.391%
UPSTREAM	1,710,496	9.508%
UTR_3_PRIME	99,605	0.554%
UTR_5_PRIME	28,764	0.16%

Comparing GATK variants found in the SRA and DNB datasets

As detailed in the GATK method, g.VCF GATK HaplotypeCaller primary calls of both GATK analyses where combined and the resulting 2-genome VCF file was processed with GATK-VQSR, and used to compare variant calls between the two data-sets. SNPEff and SNPSift were used to annotate the VCF data and extract counts for different metrics used to summarize and plot using R.

Plot Legend:

- ploidy was removed to plot below, '0/1' or '1/1' genotypes are both considered 'variant'
- half calls are silmplified to the closest allele.
  - '0/0' or '0/.' becomes 'ref'
  - '0/1', '1/1', or '1/.' becomes 'var'

Filter:

- GATK VQSR reported the variant as PASS or marked the variant as LQ

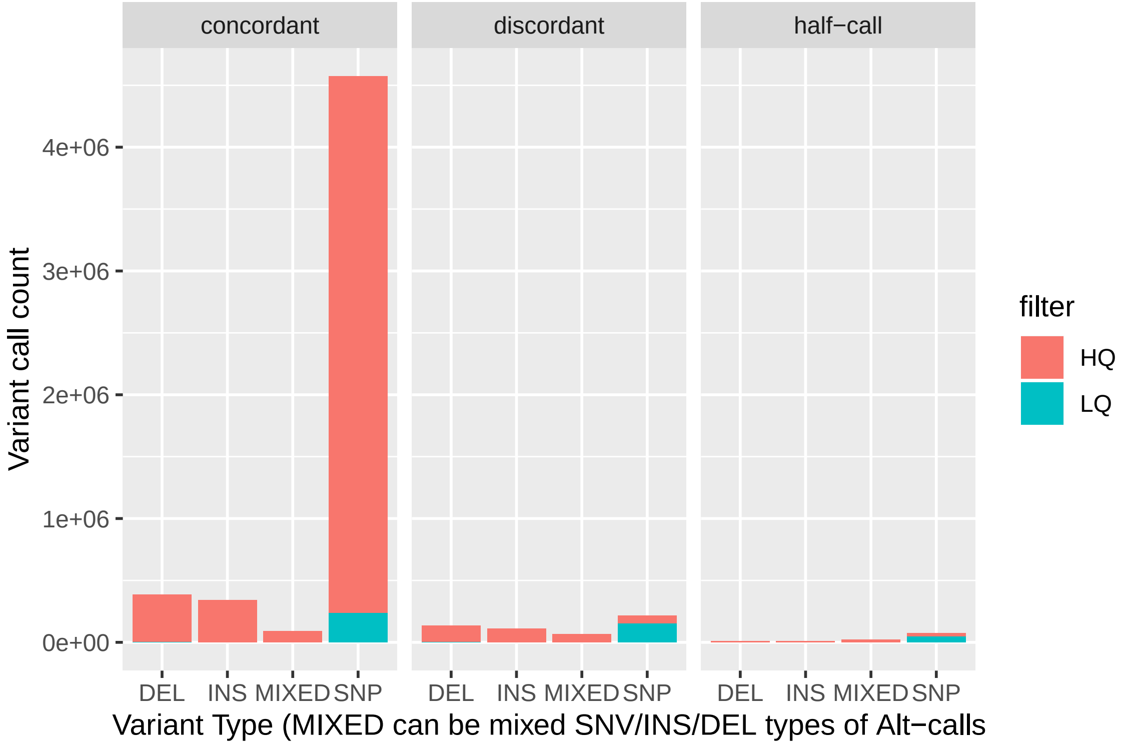
concordance:

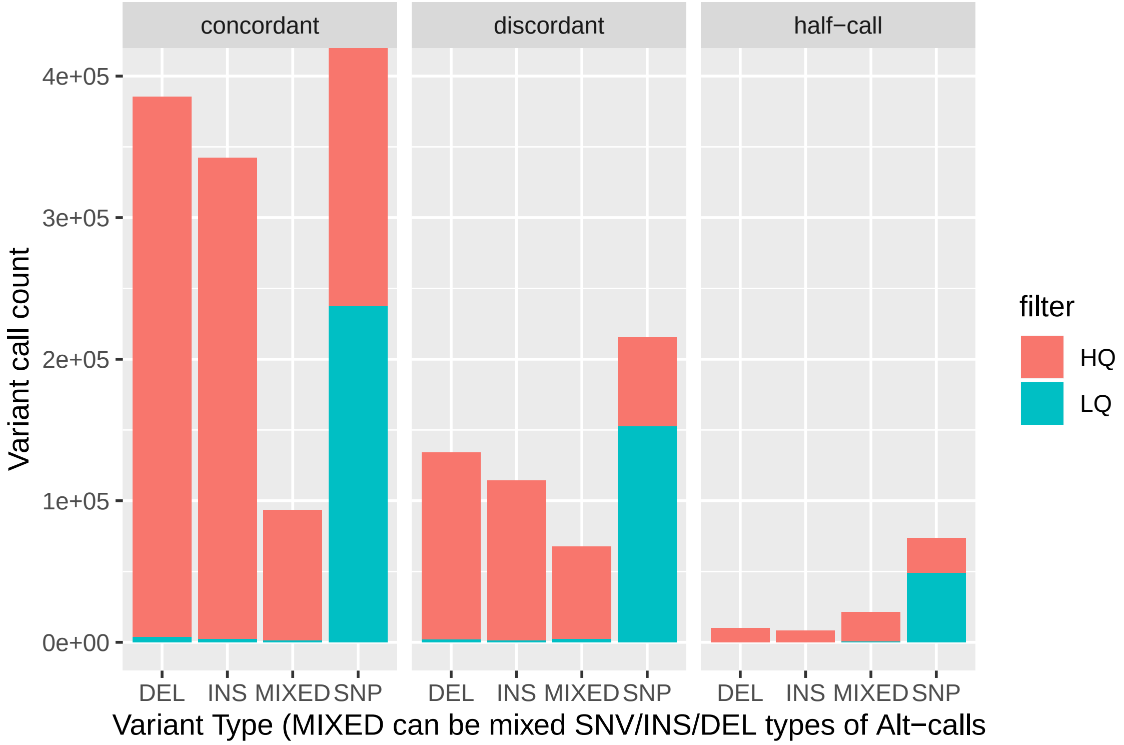
- 'concordant' means that GATK sees a 'variant' in both datasets at the same location (but can be a different variant call)
- 'discordant' means that one data say 'variant' where the other says 'reference'
- 'half-call' means that one data is 'not-called' due to lack of reads (or other technical reason)

types:

- a single SNV is called for either/both data
- a single INS (/DEL) short structural variant of the same type is called for either/both data
- MIXED can be a SNV in one data and INS or DEL in the second or more than one variant type at this position (Alt-calls)

REM: Zooming in the lower range in the second plot to see more details





Alltogether, the concordance is very good between the two data-sets. The number of 'half-calls' is quite low, confirming that both Illumina and DNB platforms reach a similar level of genome coverage.

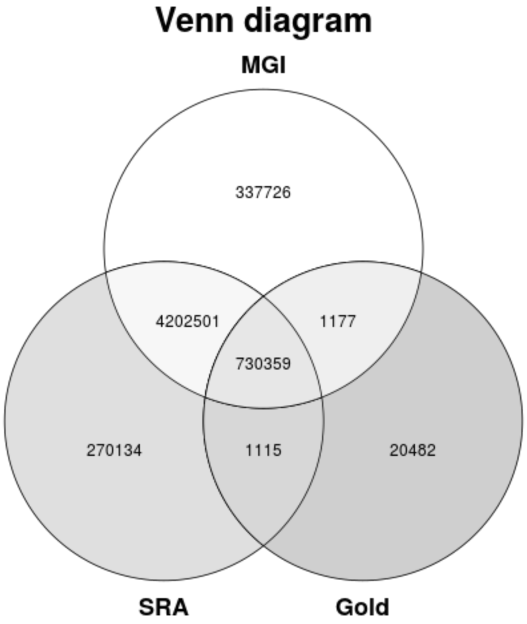
Comparing results with a Gold Standard variant set

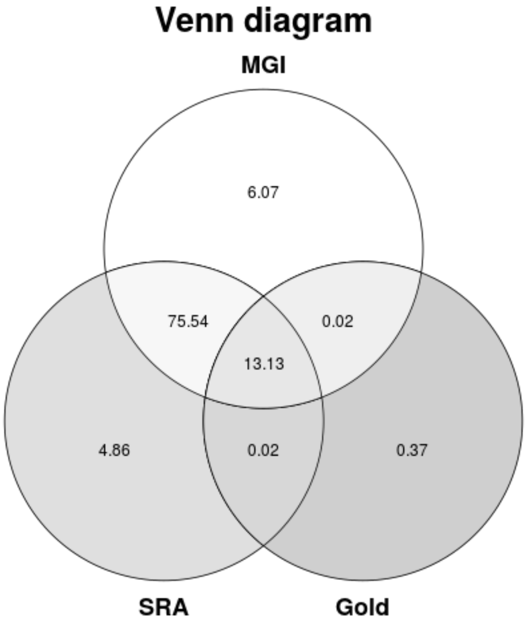
Since NA19240 was sequence multiple times including in the GIAB studies ((Zook et al. 2014), (Hwang et al. 2019)), we could retrieve gold-standard variants for that genome to be compared to our new calls.

We chose to use a 1000g call-set that results from the intersection of at least two methods and is considered as strongly enriched in true calls. The NA19240_b37 data was obtained from the UMich ftp server and lifted to the hg38 reference genome using command-line tools.

Comparing the two VCF files (SRA data and MGI data and the gold standard) was done using Vcf-Tools (vcf-compare) and the obtained counts were used to plot venn diagrams using custom code.

```
#VN 'Venn-Diagram Numbers'. Use `grep ^VN | cut -f 2-` to extract this part.
#VN The columns are:
#VN        1  .. number of sites unique to this particular combination of files
#VN        2- .. combination of files and space-separated number, a fraction of sites in the file
VN      1115    snpeff_SRR505888/NA19240_SRR505888_VQSR_snpeff.vcf.gz (0.0%)
                reference/Omni25_genotypes_NA19240_hg38.vcf.gz (0.1%)
VN      1177    snpeff_V300043186.4/NA19240_MGI_VQSR_snpeff.vcf.gz (0.0%)
                reference/Omni25_genotypes_NA19240_hg38.vcf.gz (0.2%)
VN      20482   reference/Omni25_genotypes_NA19240_hg38.vcf.gz (2.7%)
VN      270134  snpeff_SRR505888/NA19240_SRR505888_VQSR_snpeff.vcf.gz (5.2%)
VN      337726  snpeff_V300043186.4/NA19240_MGI_VQSR_snpeff.vcf.gz (6.4%)
VN      730359  snpeff_SRR505888/NA19240_SRR505888_VQSR_snpeff.vcf.gz (14.0%)
                snpeff_V300043186.4/NA19240_MGI_VQSR_snpeff.vcf.gz (13.9%)
                reference/Omni25_genotypes_NA19240_hg38.vcf.gz (97.0%)
VN      4202501 snpeff_SRR505888/NA19240_SRR505888_VQSR_snpeff.vcf.gz (80.8%)
                snpeff_V300043186.4/NA19240_MGI_VQSR_snpeff.vcf.gz (79.7%)
#SN Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN      Number of REF matches:  730299
SN      Number of ALT matches:  728904
SN      Number of REF mismatches:       60
SN      Number of ALT mismatches:       1395
SN      Number of samples in GT comparison:     0
# Number of sites lost due to grouping (e.g. duplicate sites): lost, %lost, read, reported, file
SN      Number of lost sites:   2       0.0%    753135  753133  reference/Omni25_genotypes_NA19240_hg38.vcf.gz
```





As always seen with variant analysis, a large majority of the calls found in either read types are not present in the gold-standard (~80%). By contrast the two read-types lead to very similar call sets (~75% shared). Finally, the large majority (~97%) of the Gold-standard variants is found back in both GATK sets, thereby demonstrating that the MGI reads are as good as Illumina reads, even when the later were used with a 40% excess as in this study.

Discussion & Conclusion

This supports our conclusion that MGI reads is very comparable if not of better quality than the classical Illumina reads. All QC performed in the current project resulted in 'normal' plots and counts and we could not identify biases nor issues with the MGI data.

The only remark we can add is that the insert size observed in our MGI library was rather small (~220bps) and that future genome analysis projects should aim at a slightly larger insert size during the fragmentation step in order to better discriminate in difficult genomic regions (eg insert size >300bps like for Illumina libraries).

The data shown here supports using the DNBSeqG400 platform for WGS applications when running-cost is favorable.

last edits: Fri Mar 06, 2020

VIB NUCLEOMICS CORE more at **http://www.nucleomics.be**

References

Hwang, K. B., I. H. Lee, H. Li, D. G. Won, C. Hernandez-Ferrer, J. A. Negron, and S. W. Kong. 2019. "Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings." *Sci Rep* 9 (1): 3219.

Zook, J. M., B. Chapman, J. Wang, D. Mittelman, O. Hofmann, W. Hide, and M. Salit. 2014. "Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls." *Nat. Biotechnol.* 32 (3): 246–51.