1 Genome Assembly

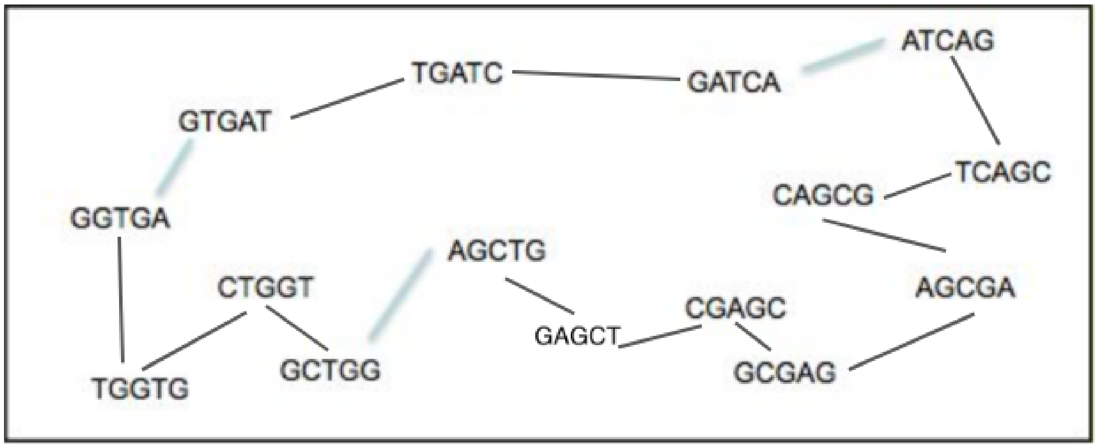
There are no questions in this section.

2 PacBio Genome Assembly

There are no questions in this section.

3 Assembly algorithms

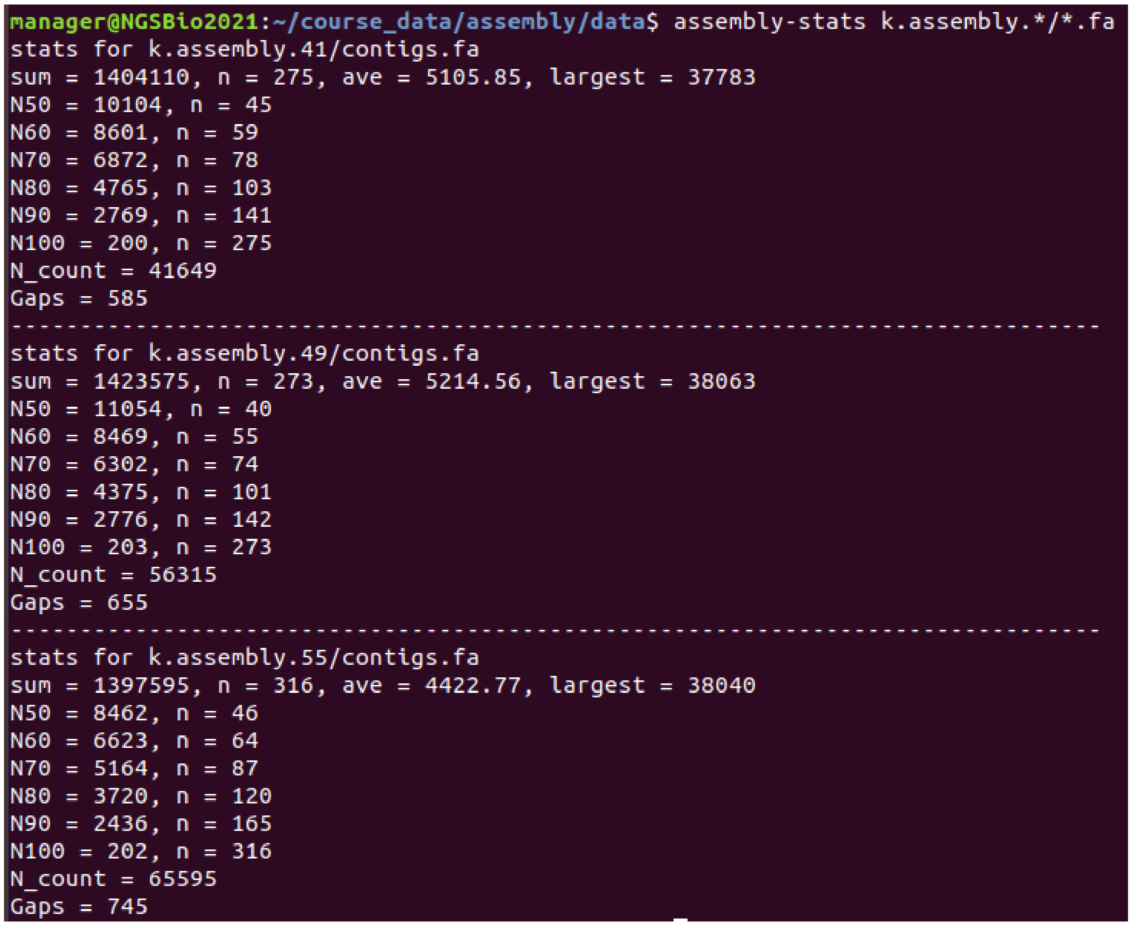
The completed graph should look like:



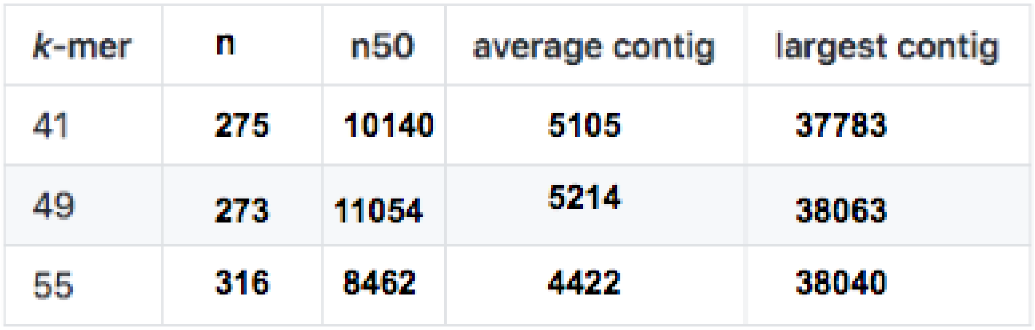
1. One possible solution is: GGTGAGTGATTGATCGATCAATCAGTCAGCCAGCGAGCGAGCGAGC-GAGCGAGCTAGCTGGCTGGCTGGTTGGTG
2. It is a circular sequence, so unable to determine the start point.

4 Illumina Genome Assembly

The results for each assembly should look something like:



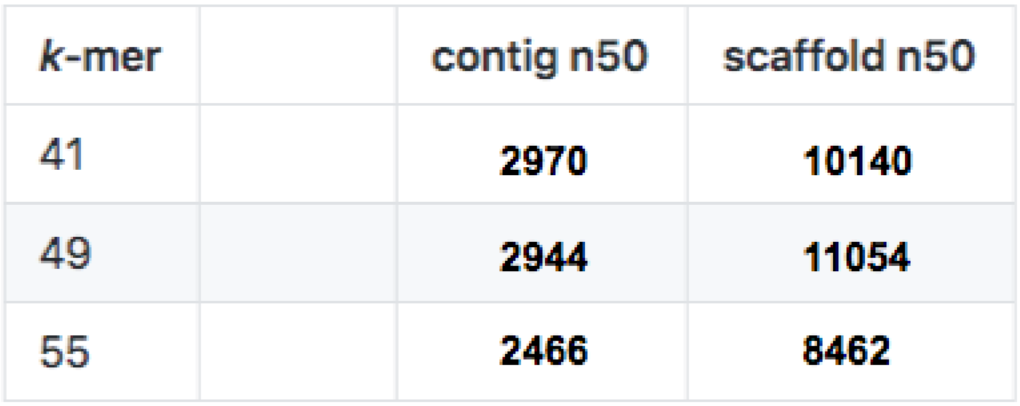
```
manager@NGSBio2021:~/course_data/assembly/data$ assembly-stats k.assembly.*/*.fa
stats for k.assembly.41/contigs.fa
sum = 1404110, n = 275, ave = 5105.85, largest = 37783
N50 = 10104, n = 45
N60 = 8601, n = 59
N70 = 6872, n = 78
N80 = 4765, n = 103
N90 = 2769, n = 141
N100 = 200, n = 275
N_count = 41649
Gaps = 585
-------------------------------------------------------------------------------
stats for k.assembly.49/contigs.fa
sum = 1423575, n = 273, ave = 5214.56, largest = 38063
N50 = 11054, n = 40
N60 = 8469, n = 55
N70 = 6302, n = 74
N80 = 4375, n = 101
N90 = 2776, n = 142
N100 = 203, n = 273
N_count = 56315
Gaps = 655
-------------------------------------------------------------------------------
stats for k.assembly.55/contigs.fa
sum = 1397595, n = 316, ave = 4422.77, largest = 38040
N50 = 8462, n = 46
N60 = 6623, n = 64
N70 = 5164, n = 87
N80 = 3720, n = 120
N90 = 2436, n = 165
N100 = 202, n = 316
N_count = 65595
Gaps = 745
```



k-mer	n	n50	average contig	largest contig
41	275	10140	5105	37783
49	273	11054	5214	38063
55	316	8462	4422	38040

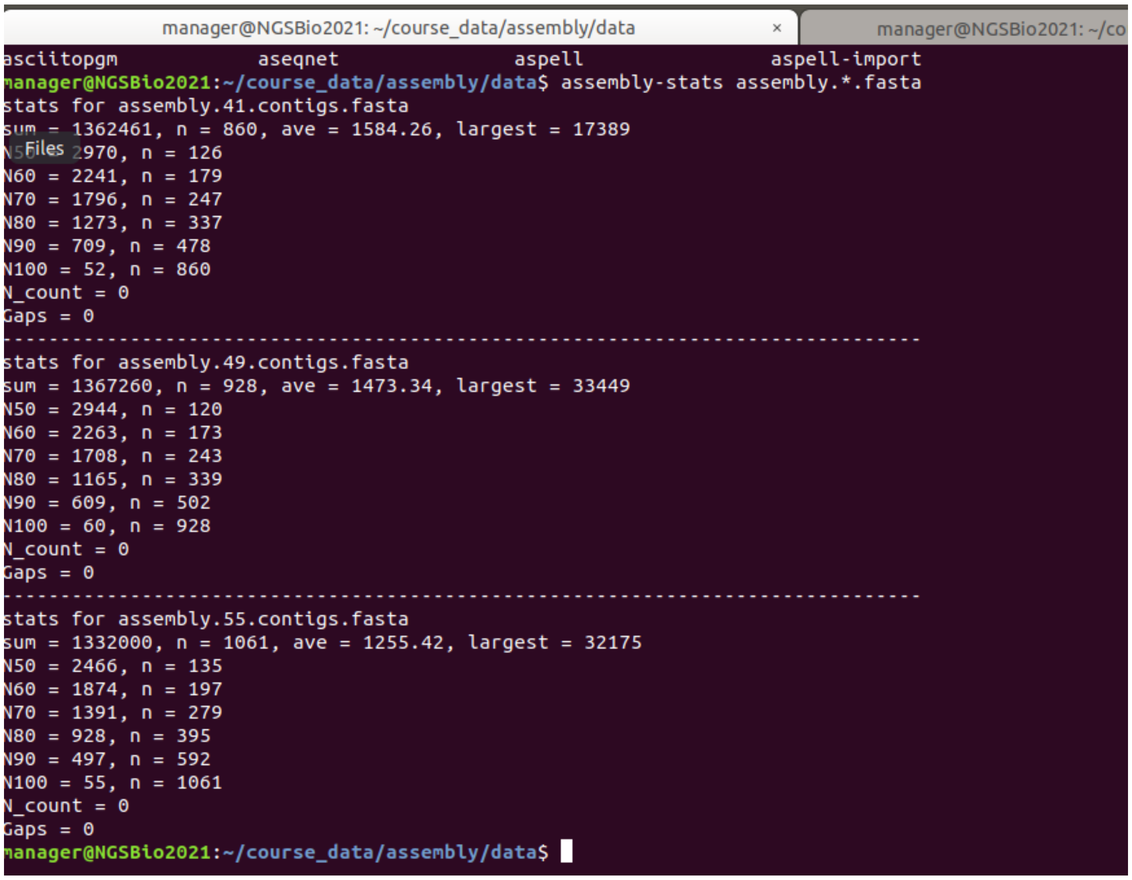
The best assembly is produced using a kmer of 49.

The statistics for the contigs versus the scaffolds should look something like:



k-mer		contig n50	scaffold n50
41		**2970**	**10140**
49		**2944**	**11054**
55		**2466**	**8462**

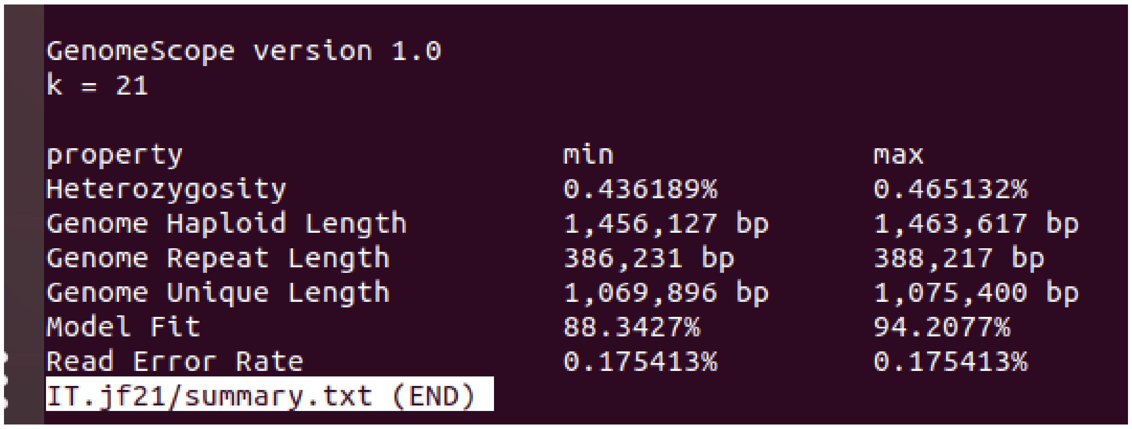
This is based on running assembly-stats for the velvet contigs:



```
manager@NGSBio2021: ~/course_data/assembly/data
asciitopgm            aseqnet               aspell                aspell-import
nanager@NGSBio2021:~/course_data/assembly/data$ assembly-stats assembly.*.fasta
stats for assembly.41.contigs.fasta
sum = 1362461, n = 860, ave = 1584.26, largest = 17389
N50 = 2970, n = 126
N60 = 2241, n = 179
N70 = 1796, n = 247
N80 = 1273, n = 337
N90 = 709, n = 478
N100 = 52, n = 860
N_count = 0
Gaps = 0
-------------------------------------------------------------------------------
stats for assembly.49.contigs.fasta
sum = 1367260, n = 928, ave = 1473.34, largest = 33449
N50 = 2944, n = 120
N60 = 2263, n = 173
N70 = 1708, n = 243
N80 = 1165, n = 339
N90 = 609, n = 502
N100 = 60, n = 928
N_count = 0
Gaps = 0
-------------------------------------------------------------------------------
stats for assembly.55.contigs.fasta
sum = 1332000, n = 1061, ave = 1255.42, largest = 32175
N50 = 2466, n = 135
N60 = 1874, n = 197
N70 = 1391, n = 279
N80 = 928, n = 395
N90 = 497, n = 592
N100 = 55, n = 1061
N_count = 0
Gaps = 0
nanager@NGSBio2021:~/course_data/assembly/data$
```

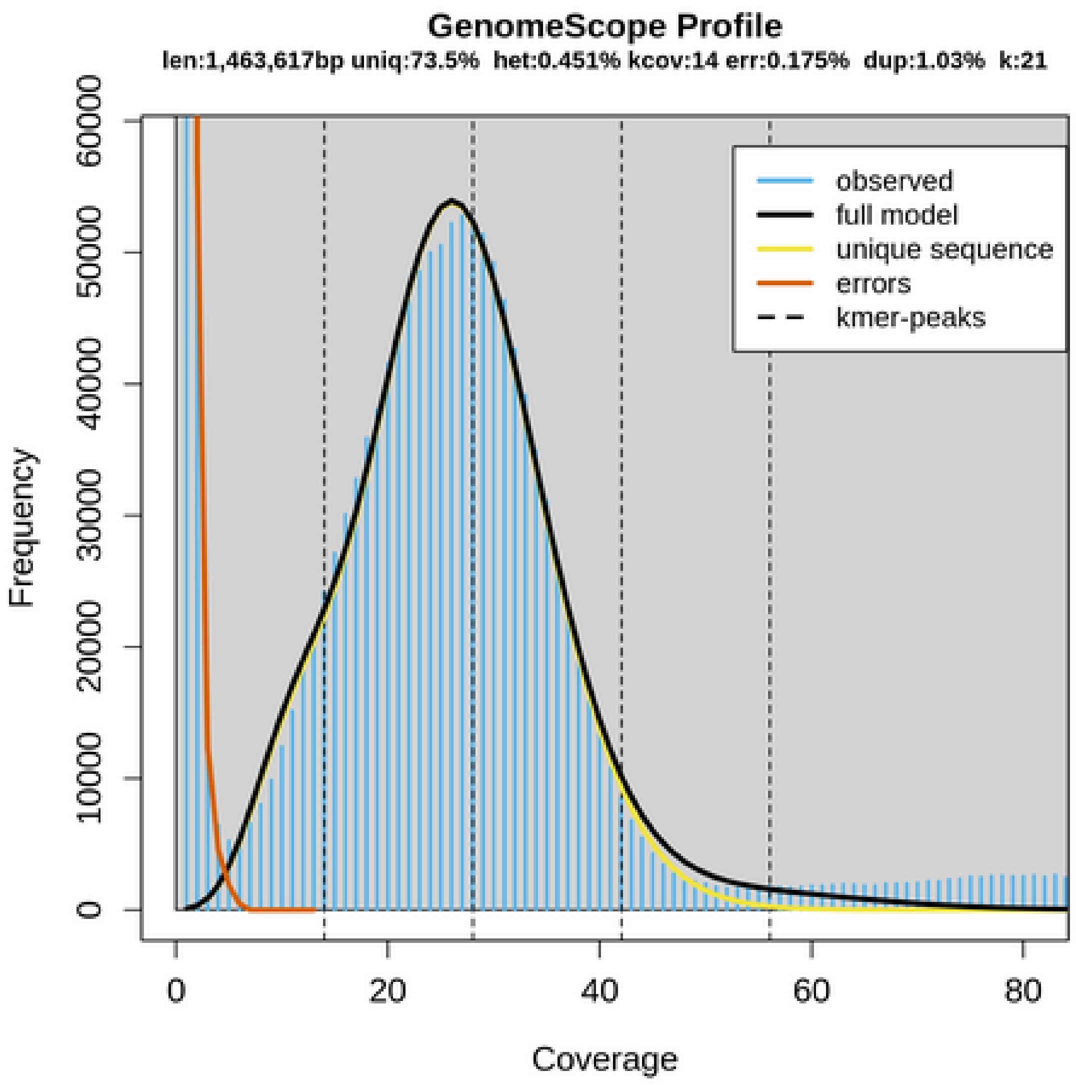
5 Assembly estimation

The output should look like:



```
GenomeScope version 1.0
k = 21

property                      min               max
Heterozygosity                0.436189%         0.465132%
Genome Haploid Length         1,456,127 bp      1,463,617 bp
Genome Repeat Length          386,231 bp        388,217 bp
Genome Unique Length          1,069,896 bp      1,075,400 bp
Model Fit                     88.3427%          94.2077%
Read Error Rate               0.175413%         0.175413%
IT.jf21/summary.txt (END)
```



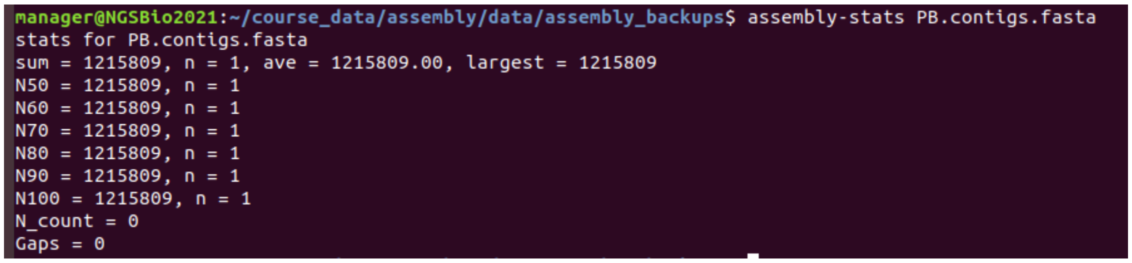
1. 0.451%
2. 1.46 MBp

3. Yes. The size of chromosome 5 of P. falciparum IT is ~1.435MBp.

6 Pacbio Genome Assembly contd.

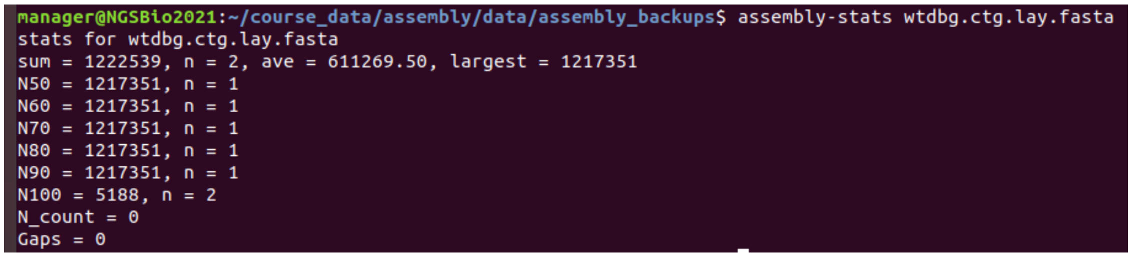
6.1 Generating pacbio assemblies

The assembly stats for the canu pacbio assembly look like:



```
manager@NGSBio2021:~/course_data/assembly/data/assembly_backups$ assembly-stats PB.contigs.fasta
stats for PB.contigs.fasta
sum = 1215809, n = 1, ave = 1215809.00, largest = 1215809
N50 = 1215809, n = 1
N60 = 1215809, n = 1
N70 = 1215809, n = 1
N80 = 1215809, n = 1
N90 = 1215809, n = 1
N100 = 1215809, n = 1
N_count = 0
Gaps = 0
```

The assembly stats for the wtbg2 pacbio assembly look like:

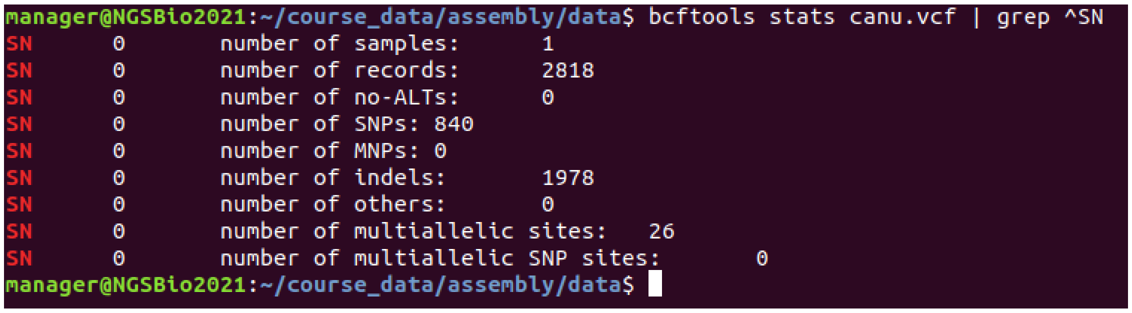


```
manager@NGSBio2021:~/course_data/assembly/data/assembly_backups$ assembly-stats wtdbg.ctg.lay.fasta
stats for wtdbg.ctg.lay.fasta
sum = 1222539, n = 2, ave = 611269.50, largest = 1217351
N50 = 1217351, n = 1
N60 = 1217351, n = 1
N70 = 1217351, n = 1
N80 = 1217351, n = 1
N90 = 1217351, n = 1
N100 = 5188, n = 2
N_count = 0
Gaps = 0
```

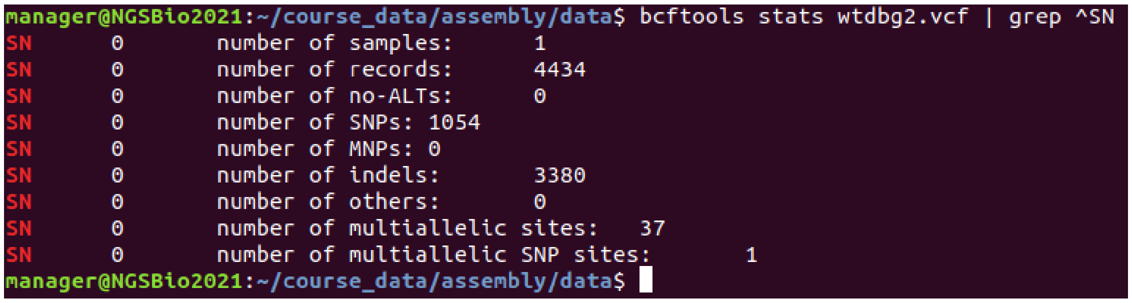
6.2 Comparing pacbio assemblies

Both assemblies are similar in terms of the common statistics (sum, N50, largest contig) with the main difference being canu assembled the sequence into one piece and wtdbg into two pieces. But let us investigate further.

The results from the variant analysis look like this, as you can see the wtdbg2 assembly has a higher number of variants.



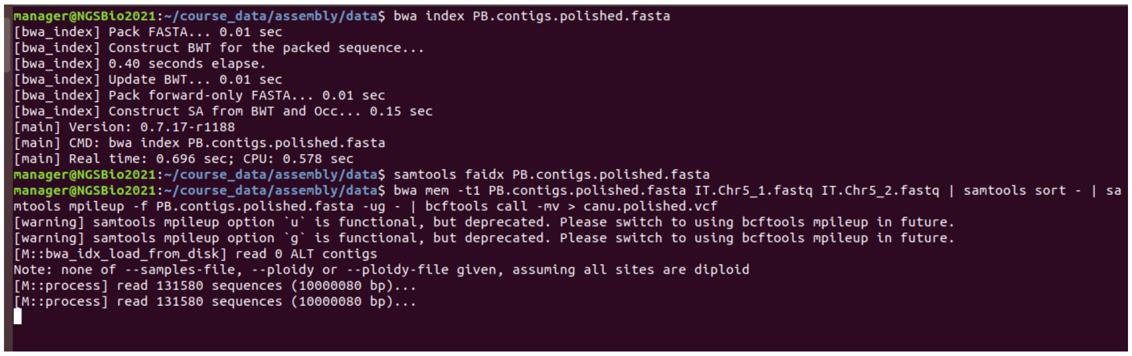
```
manager@NGSBio2021:~/course_data/assembly/data$ bcftools stats canu.vcf | grep ^SN
SN      0       number of samples:      1
SN      0       number of records:      2818
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 840
SN      0       number of MNPs: 0
SN      0       number of indels:       1978
SN      0       number of others:       0
SN      0       number of multiallelic sites:   26
SN      0       number of multiallelic SNP sites:       0
manager@NGSBio2021:~/course_data/assembly/data$ 
```



```
manager@NGSBio2021:~/course_data/assembly/data$ bcftools stats wtdbg2.vcf | grep ^SN
SN      0       number of samples:      1
SN      0       number of records:      4434
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 1054
SN      0       number of MNPs: 0
SN      0       number of indels:       3380
SN      0       number of others:       0
SN      0       number of multiallelic sites:   37
SN      0       number of multiallelic SNP sites:       1
manager@NGSBio2021:~/course_data/assembly/data$
```

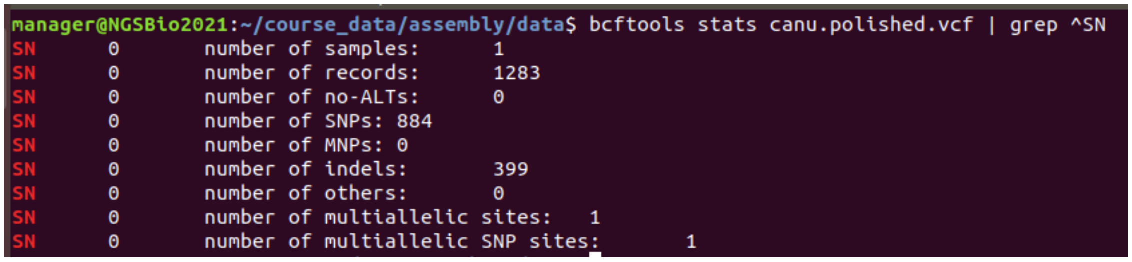
6.3 Polishing pacbio assembly

To align the reads and call variants against the polished assemblies the commands look similar to:

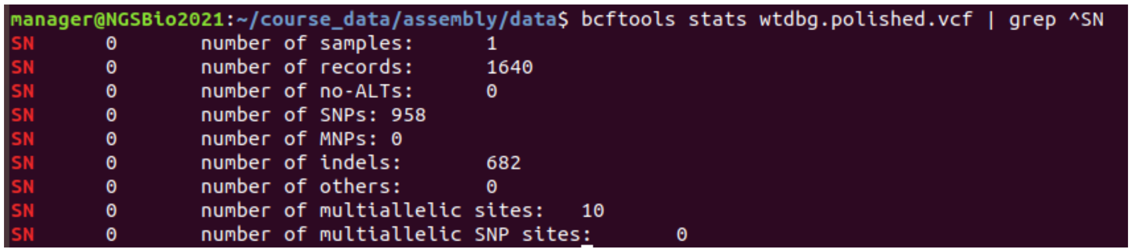


```
manager@NGSBio2021:~/course_data/assembly/data$ bwa index PB.contigs.polished.fasta
[bwa_index] Pack FASTA... 0.01 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.40 seconds elapse.
[bwa_index] Update BWT... 0.01 sec
[bwa_index] Pack forward-only FASTA... 0.01 sec
[bwa_index] Construct SA from BWT and Occ... 0.15 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index PB.contigs.polished.fasta
[main] Real time: 0.696 sec; CPU: 0.578 sec
manager@NGSBio2021:~/course_data/assembly/data$ samtools faidx PB.contigs.polished.fasta
manager@NGSBio2021:~/course_data/assembly/data$ bwa mem -t1 PB.contigs.polished.fasta IT.Chr5_1.fastq IT.Chr5_2.fastq | samtools sort - | sa
mtools mpileup -f PB.contigs.polished.fasta -ug - | bcftools call -mv > canu.polished.vcf
[warning] samtools mpileup option `u` is functional, but deprecated. Please switch to using bcftools mpileup in future.
[warning] samtools mpileup option `g` is functional, but deprecated. Please switch to using bcftools mpileup in future.
[M::bwa_idx_load_from_disk] read 0 ALT contigs
Note: none of --samples-file, --ploidy or --ploidy-file given, assuming all sites are diploid
[M::process] read 131580 sequences (10000080 bp)...
[M::process] read 131580 sequences (10000080 bp)...
```

The results from the variant analysis on the polished assemblies look like this:



```
manager@NGSBio2021:~/course_data/assembly/data$ bcftools stats canu.polished.vcf | grep ^SN
SN      0       number of samples:      1
SN      0       number of records:      1283
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 884
SN      0       number of MNPs: 0
SN      0       number of indels:       399
SN      0       number of others:       0
SN      0       number of multiallelic sites:   1
SN      0       number of multiallelic SNP sites:       1
```



```
manager@NGSBio2021:~/course_data/assembly/data$ bcftools stats wtdbg.polished.vcf | grep ^SN
SN      0       number of samples:      1
SN      0       number of records:      1640
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 958
SN      0       number of MNPs: 0
SN      0       number of indels:       682
SN      0       number of others:       0
SN      0       number of multiallelic sites:   10
SN      0       number of multiallelic SNP sites:       0
```

As you can see you get less variants called for the polished assemblies, this is because the errors in the assembly have been corrected.