## I. Background, goals and significance

The cultivated potato (*Solanum tumberosum* L., 2n=4x=48) was domesticated in South America around 8000 years ago and is now the third most important food crop in the world (http://www.fao.org/faostat/en/#rankings/commodities_by_region). It is clonally propagated crop and has a complex genome which results in a number of challenges for genome assembly and annotation. It is an autotetraploid, meaning it's polyploidy arose from a whole genome duplication. This is associated with highly repetitive genomic regions, which makes resolving subgenomes difficult. It is also an obligate outcrosser, an evolutionary mechanism designed to promote genetic diversity, which has led to the maintenance of a highly heterozygous genome even after thousands of years after domestication (The Potato Genome Sequencing Consortium). Increase in heterozygosity makes it difficult to differentiate whether alleles should be mapped to the same loci, or whether they are different members of the same gene family (Huang et al, 2017). Advancements in technology and computational methods move to resolve the limitations associated with complex polyploid genomes.

A reference genome for potato was developed in 2011 by sequencing a doubled monoploid of *Solanum tuberosum* Group Phruja species. The use of a homozygous doubled monoploid simplified the assembly of the reference which resulted in 86% of the 844 megabase genome being assembled using next generation sequencing platforms. To assess whether this homozygous 'diploid' genome could handle the heterozygosity of cultivated potato, RNA-Seq data was generated from the RH cultivar. 88.6% of the reads mapped using the aligner Tophat, validating the ability to map most of the reads even with genetic complexity (The Potato Genome Sequencing Consortium, 2011).

The availability of a high quality reference genome means that further exploration of the potato genome can be done. RNA-seq is a high throughput sequencing method that is employed to measure expression of genes. RNA-Seq is often employed for differential expression studies where RNA is extracted, converted to cDNA, sheared into small fragments and then mapped to a genome. The abundance of reads is used to measure the level of expression. The short reads can be mapped to a reference genome or a transcriptome can be generated *de novo* if a reference is unavailable (Costa- Silva et al, 2017).

There are a number of alignment algorithms that are available for mapping RNA-seq reads, but new aligners and updates to previous aligners have made read mapping more efficient as technologies have also advanced. I am interested in comparing 4 different aligners which can be classified as either splice aware or quasi-alignment approaches. Splice aware aligners (STAR, HISAT2, and SOAPSplice) acknowledge that when aligning RNA reads to a reference genome that reads will not span introns present in the reference. Quasi-alignment approaches like Salmon skip the step of getting a full alignment of the reads to the transcriptome, and instead uses Maximal Exact Matches (MEM) to estimate transcript abundance.

The complexity of the potato genome has been a huge limitation to breeding efforts, as progress from genetic advances has not made significant changes in yield in the past century (Bradshaw, 2017). Therefore, unraveling this complexity through genomic tools can help overcome these limitations. This kind of comprehensive analysis of read aligners will create a pipeline for future bioinformatic studies in cultivated potato. This is a useful resource for the future of potato breeding and can also be applied to a number of genomically complex plant systems.

## II. Dataset

The RNA-seq data used for this study was provided by Dr. Robin Buell. RNA was extracted from the leaf tissue from the elite commercial cultivar Missaukee. Libraries were prepared by Novogene and ran on Illumina HiSeq 4000 to generate 32,625,431, 150bp paired end reads (Pham et al, 2017). Samples were barcoded using NEBnext Multiplex Oligos for Illumina.

.

## III. Computational Methods

*Alignment algorithms*

Splice aware aligners take into account the boundaries between exon and introns when mapping reads to a reference. They can either use annotation of gene structures, or the algorithm itself can identify splice junctions. The splice aware aligners used in this study will be STAR, HISAT2,  and TopHat2.

STAR (Splice Transcripts Alignment to a Reference) is capable of aligning non-contiguous short read sequences to a reference genome. Similar to MEM strategies, STAR uses a Maximal Mappable Prefix (MMP) to map reads and  using uncompressed suffix arrays to identify how the reads align (Figure 1). It starts at the beginning of a read and finds the longest contiguous sequence that aligns to the reference (MMP1), it then repeats this process to search for the acceptor site with the rest of that read sequence (MMP2). Each portion of the read is referred to as a seed. This method is much faster than aligners like Mummer because it does not find all MEMs, only the ones within a certain distance of MMP1. After identifying all of the MMPs, it clusters local alignments together to resolve gaps.  They are then scored based on number of mismatches and indels. The seeds with the highest scores are the ones used for the alignment. (Dobin et al, 2013).

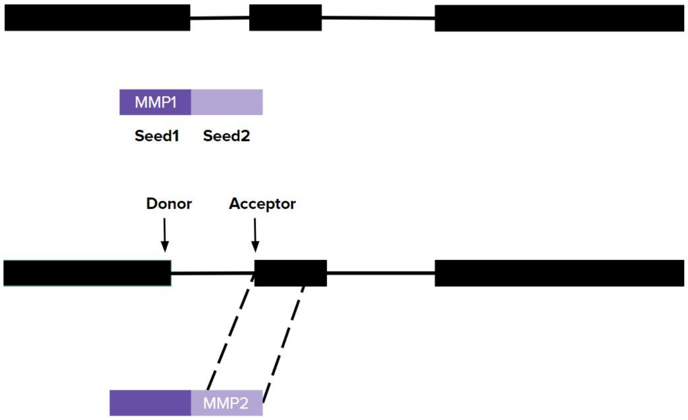


Figure 1. Illustration of how STAR maps reads across splice junctions

TopHat was one of the first softwares available that could predict possible splice junctions in neighboring exons, thus being useful for RNA sequencing. TopHat2 maps reads to the reference using Bowtie. This relies on Burrows-Wheeler transformation and the Ferragina-Mazini (FM) index to index alignments. Reads that can map to the reference are used to identify potential exons and splice junctions. The reads that were initially unmapped are then mapped to these predicted junctions.

HISAT2 (Hierarchical indexing for spliced alignment of transcripts) was developed by the creators of TopHat to increase the speed in spliced read alignment. It also maps reads to the reference using Bowtie, creating a whole genome FM index and then creating local FM indices to extend the alignments. The benefit of HISAT2 is the utility of using local indexes to anchor reads instead of just global alignment like TopHat. Similar to STAR, it first maps the first segment of a read and then identifies where the second segment maps to (Kim et al, 2015).

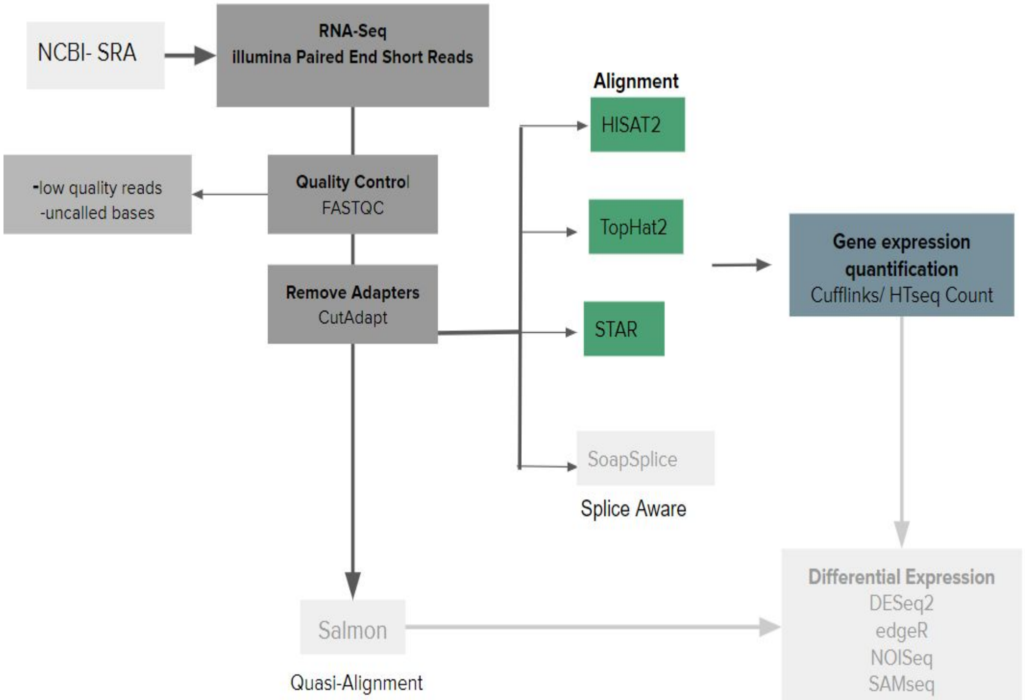


**Figure 2.** Outline of RNA-Seq alignment pipelines. Light gray boxes correspond to parts of the original pipeline that weren't addressed in final report.

## IV. Evaluation

*Alignment algorithms*

To evaluate the how well the aligners works, a suitable metric will be comparing the number of raw reads that are converted to countable reads. For the splice- aware aligners, counting the number mapped reads will be presented in their outputs. A failure to convert raw reads can be due low quality read alignment, multiple alignments, or missing reference annotation, all of which varies depending on the aligner chosen.  The number of countable reads are an important metric because a high read depth increases the sensitivity of differential gene expression analysis. Only reads that map to exons are counted for transcript abundance so itt may also be of interest to see where the reads are being mapped to, whether it be exonic or non-exonic (intergenic) regions (Paya-Milans et al, 2018). The speed of each aligners will also be taken into account as a measure of how user friendly these softwares can be.

The paired end reads from one biological replicate from leaf tissue were mapped using each aligner with their default settings to the potato reference genome (V4.03) and using the genome annotation (V4.04). All work was conducted in a high performance computing cluster

## V. Results

RNA-Seq reads were first quality checked using FastQC reports (Figure 3). FastQC reports summarize the raw data so that the user can identify problems or biases in the read quality that can affect downstream analysis. Phred quality scores (Figure 3A) are a measure of confidence in the quality of each base generated from a next generation sequencing platform. The scores are logarithmically related to the probability that a base was called incorrectly. A Phred score of 40 would represent a 1 in 10,000 chance that there was an error in the base that was called. The error rate is assessed based on the the shape and resolution of the fluorescent peaks generated by the sequencer. The reads from this dataset show high quality reads as the average score is 38, showing confidence in the sequences generated in this data. The bases per sequence content should reflect the average base content in the genome. The potato genome does not have over represented bases therefore it is expected that each base would exist in equal proportions in the reads (Figure 3B). Biased fragmentation caused the imbalance at the beginning of the reads in this dataset as is typical in RNA-seq libraries. This is due to the bias of certain “random” hexamers tagging sequences during cDNA library preparation. The GC content (Figure 3C) is expected in *most* genomes to be a normal distribution, a deviation from this may indicate that contamination in the library. The first 100,000 sequences in the dataset are used by FastQC to identify sequence duplications (Figure 3D). Overrepresented sequences can indicate PCR over amplification, that is,  a certain section of the sequencing lane was producing more reads than the rest. It is typical in RNA-Seq data to have over enriched sequences as these may represent genes that are more highly expressed.

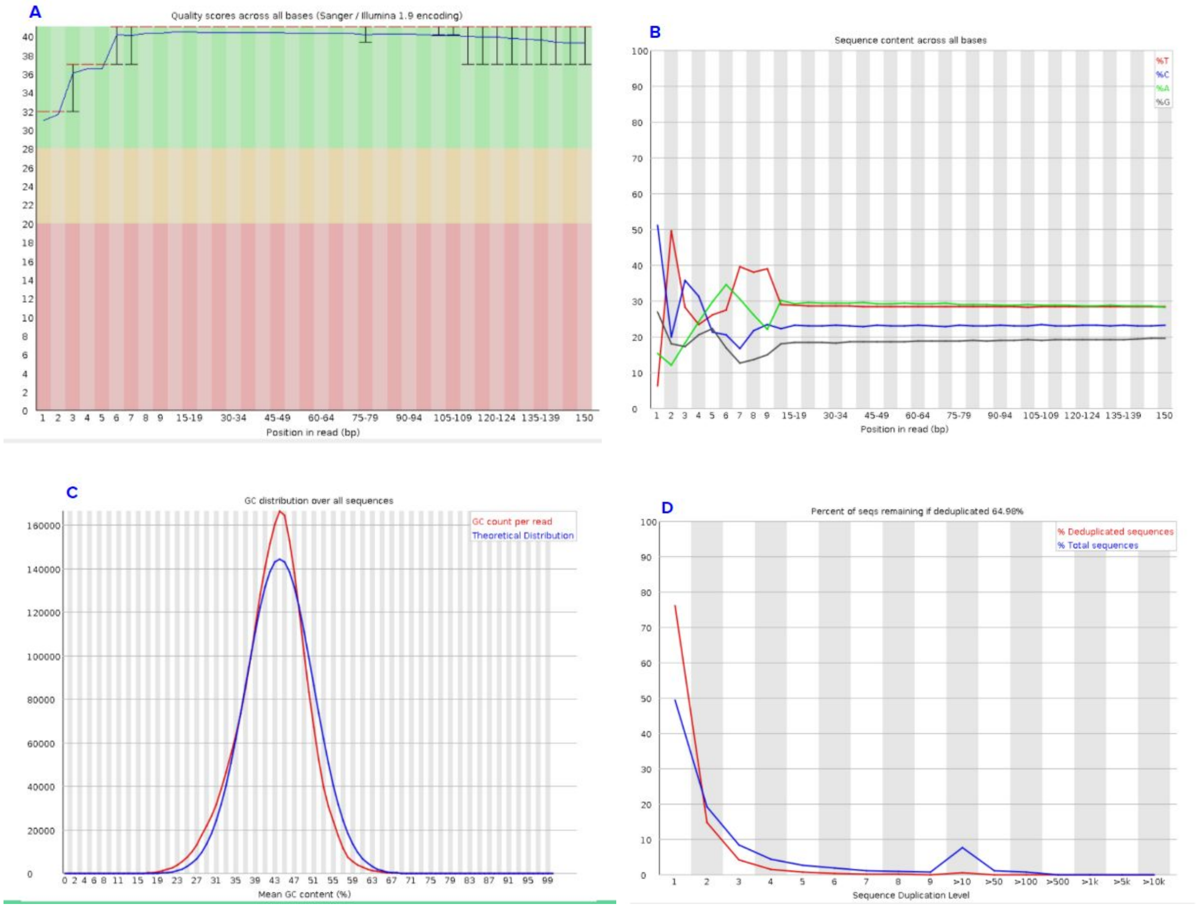


**Figure 3**. FastQC results from Missauke Leaf replicate 1. (A) Average phred quality scores for base calls, (B) per base sequence content, (C ) per sequence GC content, (D) sequence duplication levels.

Sequencing adapters were removed using Cutadapt which is a software that removes low quality reads and removes sequencing adapters from short read data sets. Adapters contain indexes from the NEBNext Multiplex oligos, therefore each biological sample had it's corresponding adapter removed. Reads were removed if their average Phred score was below 30 and/orif the read was less was less than 100 bp. It took 1,563.65 seconds to remove adapters and low quality reads from all 32, 625,431 PE reads, processing at 1.38 million reads per minute (Supplementary Fig 1).

The first step in using an aligner is to first index the reference genome.  For each aligner only the default index settings were used (Table 1.). Using uncompressed suffix arrays, STAR  is capable of indexing the genome quickly in 10 minutes and 47 seconds, but takes up a lot of memory in the process using about 17.6 GB . HISAT2 is an intermediate with taking only 8:25 minutes to index the reference using Bowtie and consumes only about 1.2 GB of memory. TopHat2 also uses Bowtie2 to index, taking 11:08 minutes and using  1.25 GB of memory.

After the genome is indexed, the reads were aligned using default settings except for TopHat2 (Table 1.). STAR was the fastest, taking only 23 minutes. HISAT2 took 1 hour and 14 minutes to run. TopHat2 took the longest to run using the default settings since it was only allowed to run on one thread as default. The default run lasted for over 15 hours, the number of parallel tasks was then increased to 4 so that the program could actually run all the way through in a reasonable amount of time. When the threads were set to 4, it took tophat 3 hours and 44 minutes to complete.

**Table 1.** Table 1 - Run time and memory usage based on mapping paired end reads for one biological replicate from leaf tissue using default parameters.

| **Aligner** | **Run Time (Index + Align)** | **Memory Usage (Index + Align)** |
|---|---|---|
| **STAR** | 00:10:47 + 00:23:17 =<br>**00:34:04 minutes** | 17.6 GB+ 7.9 GB =<br>**25.5 GB** |
| **HISAT2** | 00:08:25 +01:14:07<br>**1:22:32 hours** | 1.2 GB + 1.1 GB = **2.3 GB** |
| **TopHat2** | 00:11:08 + > 15:00:00<br>**>15 hours** | 1.25 GB + 3.4 GB= **4.65 GB** |

Using STAR to align one biological replicate from leaf tissue that had 32,625,431 raw reads to the reference genome resulted in 90.67% of reads uniquely mapping (Figure 4), with 4.95% of the reads mapped to multiple loci. Using TopHat2 to align the same reads, it had and overall 68.8% mapping rate, of which 3.5% had multiple alignments and 1.1% were discordant alignments. There was a 55.4% concordant pair alignment rate. Concordant read pairs are those that map in the expected orientation and size to each other, therefore we have more confidence in the mapping of these reads. Using HISAT2 to align the same reads, it had an overall 87.76 % mapping rate of which 77.66% mapped uniquely and 3.14% of the reads mapped to more than one loci.

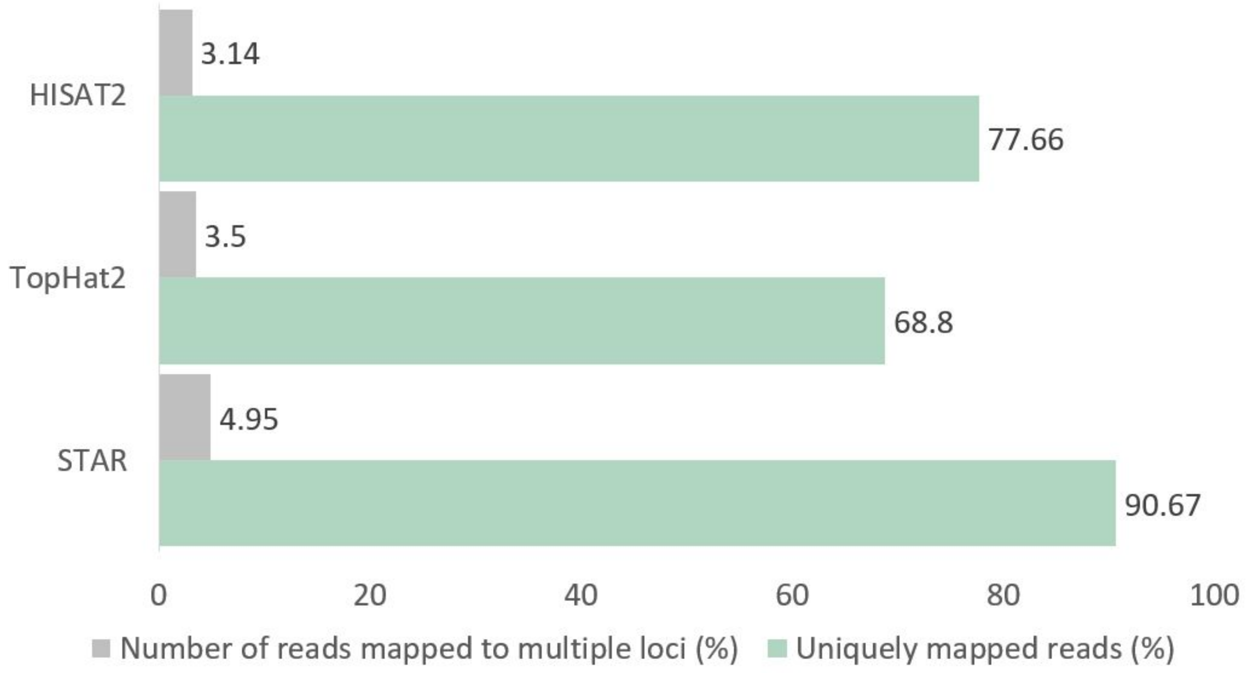


Figure 4. Percent of total reads that uniquely mapped (green) or mapped to multiple loci (gray)

## VI. Discussion

As far as ease of use, STAR was the most straight forward and gave the quickest results. When troubleshooting errors, it is much easier to wait 20 minutes compared to about 20 hours so see results. TopHat2, although once a popular aligner, was the most difficult to work with. It took the longest to run and required the most troubleshooting. It was the only aligner that threw errors on how the genome annotation was formated, and required the most data wrangling to organize the data in a way that it could actually read. About 7% of the annotations didn't have strand information and TopHat2 was the only aligner that was unable to handle this or else it would run into a parsing error while mapping reads. HISAT2 was intermediate in ease of use, it was able to handle the original genome annotation without having to delete the strandless data. It also took an intermediate amount of time to run.

STAR was able to map more reads uniquely than HISAT2 or TopHat2. This may be due to its ability to identify splice junctions more efficiently than the other two aligners. A notable difference between STAR and TopHAT/HISAT2 is on how the genome is indexed. TopHAT2 and HISAT2 both built off of Bowtie2 which was initially designed for DNA alignment and is optimized for mapping reads without intron sized gaps. Therefore it may not be as well equipped at predicting splice junctions. These limitations can be overcome in FM based aligners by optimizing alignment parameters, such as permitting higher mismatch and indel rates when aligning reads.

The significant difference in speed of STAR compared to TopHat2 and HISAT2 comes at the cost of a larger memory requirement. It does not compress it's suffix array as does FM index based arrays. Compression of indexes reduces memory requirements but makes searching for strings in the compressed matrix less efficient.

Future directions for this project would be to validate the accuracy of these aligners using simulated read data similar to that of tetraploid potato with known mapping positions. This way, you can identify the false positive rate. It would also be beneficial to have experimentally validated expression data to ensure that the read depth truly does represent what is predicted biologically.

*Project Reflection*

1.) I have a more comprehensive understanding of the entire RNA-Seq pipeline from the initial experiment set up to the actual data analysis. Along with that, I now understand how each of these softwares actually work, which makes interpreting results/anomalies easier. 2.) I feel much more confident working independently on a bioinformatics project, I still don't understand everything, but I have gotten very good at googling my problems and trying to piece together what the solution should be. 3.) This project helped me decide which aligner may be best for the data I plan on working with in the near future for my own RNA-seq experiment. I can now use this pipeline as a tool for future research, and I can go into that research with more confidence in my results. Or at least, I can better understand where more questions may arise instead of blinding accepting the outputted results of each software. 4.) Big data is pretty annoying to work with. Honestly, the majority of my time spent on this project was trying to figure out how the data should be formatted so that I wasn't getting errors. It was kind of surprising to me how files obtained from databases don't have a uniform way of being organised and labeled. Also the lack of information that comes with publicly available datasets. Luckily, the reference genome/annotation and the sequencing reads were developed by a member of my committee so it was convenient to ask questions about the data, but to anyone else, it may be difficult to navigate through.

This semester project was my first attempt at independently handling a computational project. Admittedly, it was much more difficult than I had initially thought. Reading through the software manuals makes the process seem very simple, but in reality the majority of my time was spent troubleshooting all of the errors. It was never as simple as just plugging in my files into a command. If I could start the project over, I would use a dataset that had some experimental gene quantification data available. There was not a control for this study, and false positives could not be identified. The biggest limitation to the progress of this project was the time I could dedicate to it. As I am a novice to computational work, much of my time was spent teaching myself the basics or troubleshooting errors that may have evident to someone with more experience. I will say that this method of 'trial and error' self teaching is how I learn best. It forced me to really understand the data I was working with and how seemingly little things could really throw off how a software interprets it. Although I didn't accomplish as much as someone with more experience could have done with ease, I now have much more confidence working with big data. I think getting the chance to actually apply myself instead of just learning something idly really solidified these skills for me. I thoroughly enjoyed the opportunity to play around with bioinformatics tools, with an imminent deadline that forced me to actually follow through. This was a fun class, I learned a lot. This was one of the first classes I actually felt comfortable asking questions without feeling like an idiot when I didn't understand something. Thank you for being the kind of teacher who actually invests in their students, I really appreciated it.

## References

Bradshaw, John E. 2017. Review and Analysis of Limitations in Ways to Improve Conventional Potato Breeding. *Potato Research* 60: 171–193. doi:10.1007/s11540-017-9346-z.

Costa-Silva, Juliana, Douglas Domingues, and Fabricio Martins Lopes. 2017. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* 12. doi:10.1371/journal.pone.0190152.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21. doi:10.1093/bioinformatics/bts635.

Huang, Shengfeng, Mingjing Kang, and Anlong Xu. 2017. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 33: 2577–2579. doi:10.1093/bioinformatics/btx220.

Huang, Songbo, Jinbo Zhang, Ruiqiang Li, Wenqian Zhang, Zengquan He, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. 2011. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Frontiers in Genetics* 2. doi:10.3389/fgene.2011.00046.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* 12: 357–360. doi:10.1038/nmeth.3317.

Li, Jun, and Robert Tibshirani. 2013. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research* 22: 519–536. doi:10.1177/0962280211428386.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods* 14: 417–419. doi:10.1038/nmeth.4197.

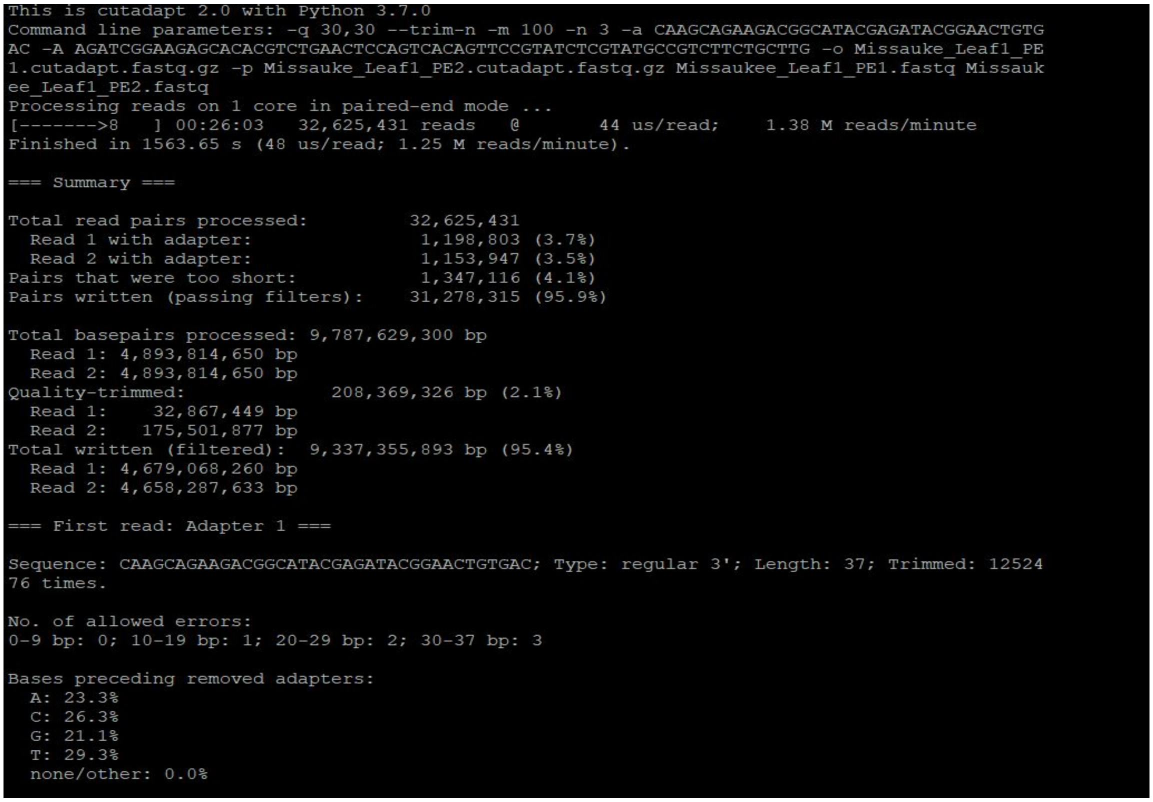
Payá-Milans, Miriam, James W. Olmstead, Gerardo Nunez, Timothy A. Rinehart, and Margaret Staton. 2018. Comprehensive evaluation of RNA-seq analysis pipelines in diploid and polyploid species. *GigaScience* 7. doi:10.1093/gigascience/giy132.

Pham, Gina M., Linsey Newton, Krystle Wiegert-Rininger, Brieanne Vaillancourt, David S. Douches, and C. Robin Buell. 2017. Extensive genome heterogeneity leads to preferential allele expression and copy number-dependent expression in cultivated potato. *The Plant Journal* 92: 624–637. doi:10.1111/tpj.13706.

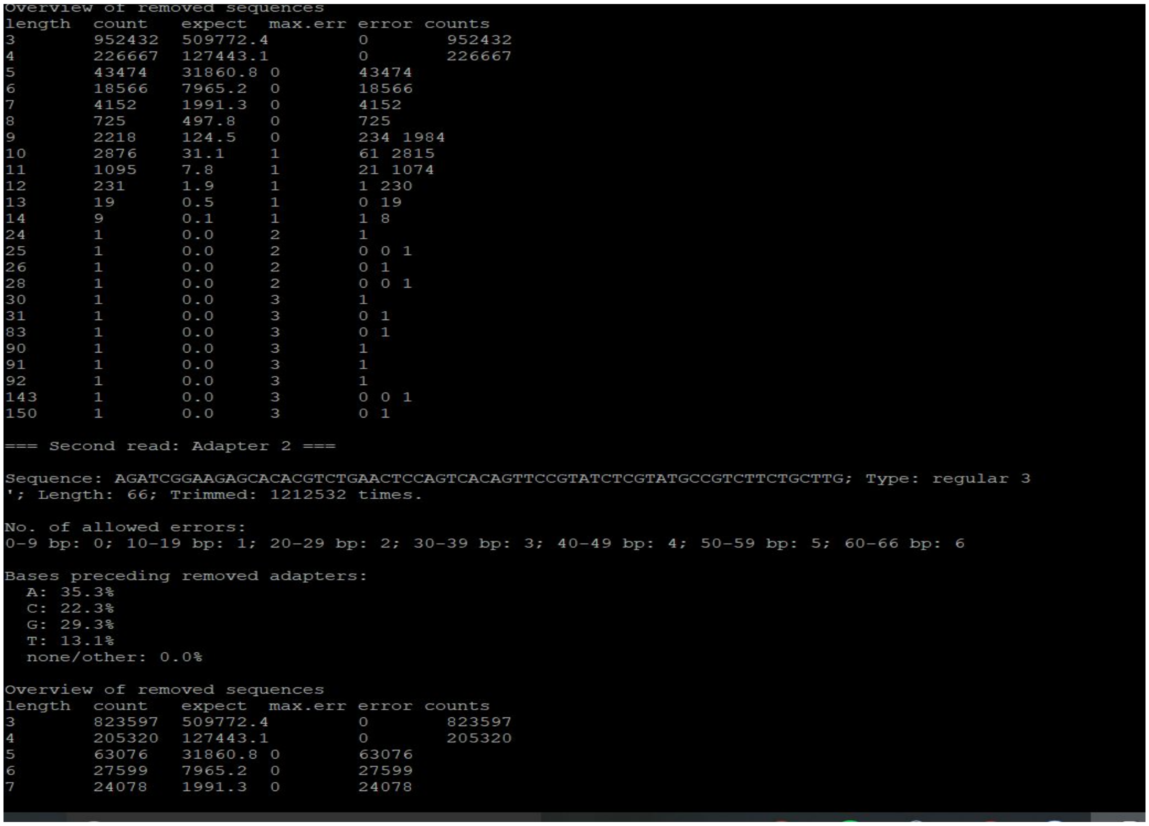
Tarazona, Sonia, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research* 43: e140–e140. doi:10.1093/nar/gkv711.

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature* 475: 189–195. doi:10.1038/nature10158.

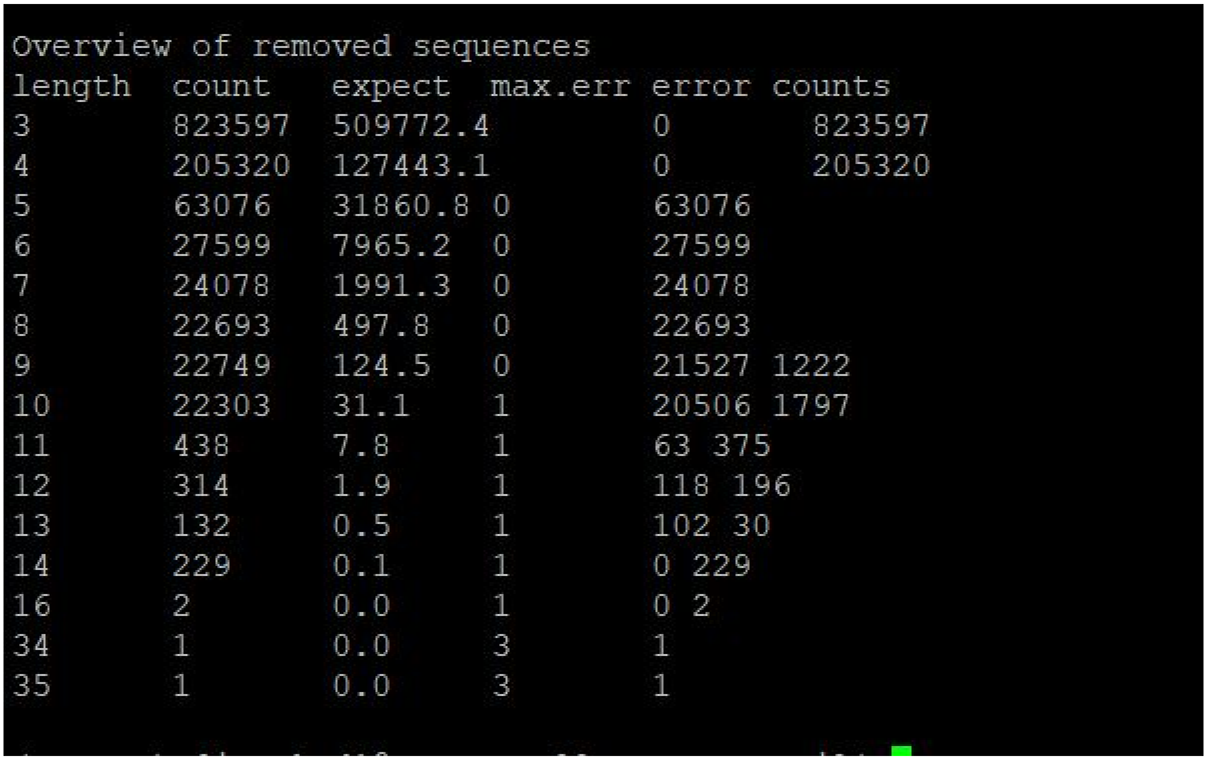
**Supplementary Figure 1.** Cutadapt output for Missaukee leaf PE reads



```
This is cutadapt 2.0 with Python 3.7.0
Command line parameters: -q 30,30 --trim-n -m 100 -n 3 -a CAAGCAGAAGACGGCATACGAGATACGGAACTGTG
AC -A AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCGTATCTCGTATGCCGTCTTCTGCTTG -o Missauke_Leaf1_PE
1.cutadapt.fastq.gz -p Missauke_Leaf1_PE2.cutadapt.fastq.gz Missaukee_Leaf1_PE1.fastq Missauk
ee_Leaf1_PE2.fastq
Processing reads on 1 core in paired-end mode ...
[------->8   ] 00:26:03   32,625,431 reads   @       44 us/read;    1.38 M reads/minute
Finished in 1563.65 s (48 us/read; 1.25 M reads/minute).

=== Summary ===

Total read pairs processed:         32,625,431
  Read 1 with adapter:               1,198,803 (3.7%)
  Read 2 with adapter:               1,153,947 (3.5%)
Pairs that were too short:           1,347,116 (4.1%)
Pairs written (passing filters):    31,278,315 (95.9%)

Total basepairs processed: 9,787,629,300 bp
  Read 1: 4,893,814,650 bp
  Read 2: 4,893,814,650 bp
Quality-trimmed:             208,369,326 bp (2.1%)
  Read 1:    32,867,449 bp
  Read 2:   175,501,877 bp
Total written (filtered):  9,337,355,893 bp (95.4%)
  Read 1: 4,679,068,260 bp
  Read 2: 4,658,287,633 bp

=== First read: Adapter 1 ===

Sequence: CAAGCAGAAGACGGCATACGAGATACGGAACTGTGAC; Type: regular 3'; Length: 37; Trimmed: 12524
76 times.

No. of allowed errors:
0-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-37 bp: 3

Bases preceding removed adapters:
  A: 23.3%
  C: 26.3%
  G: 21.1%
  T: 29.3%
  none/other: 0.0%
```

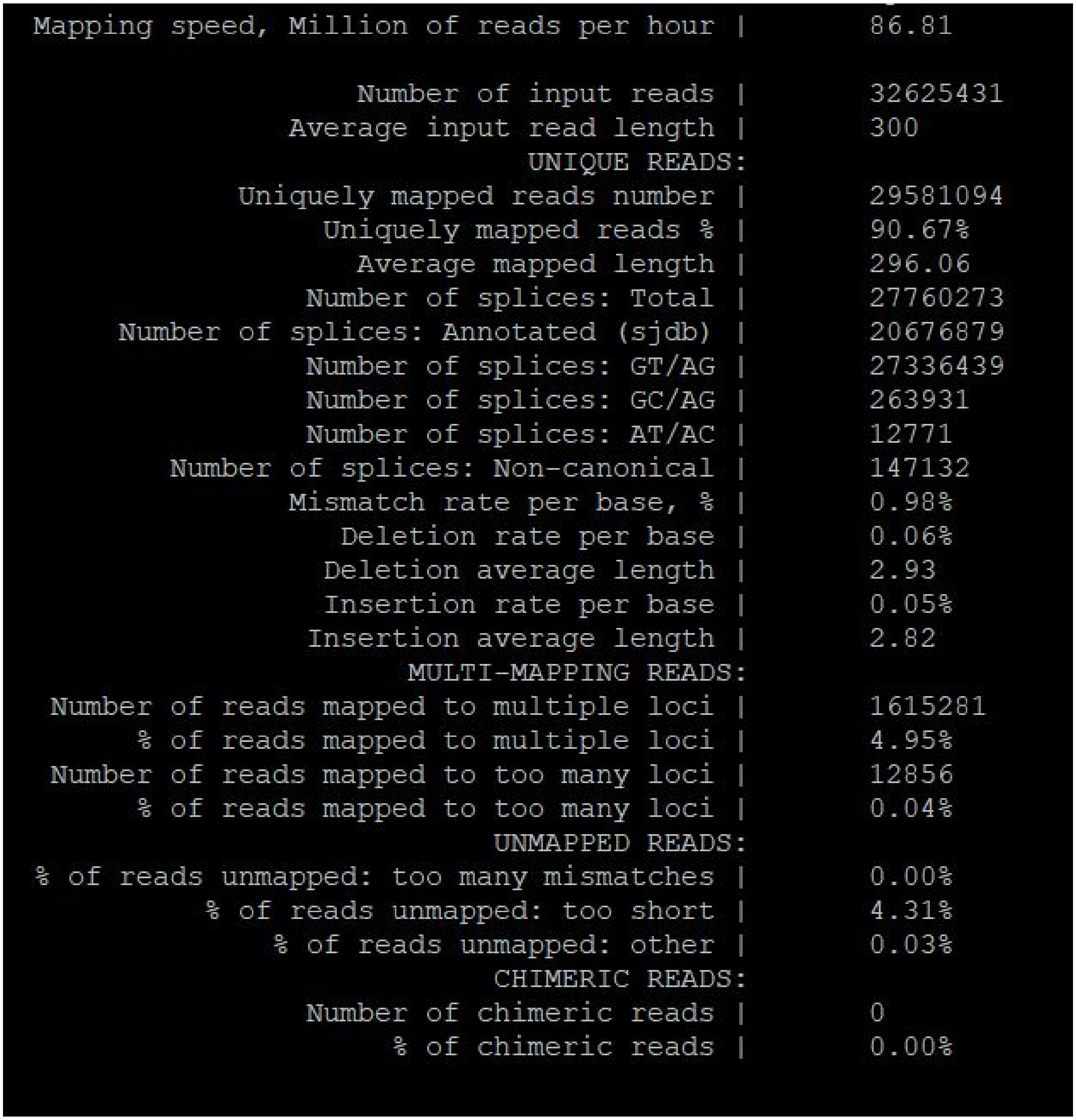


```
Overview of removed sequences
length  count   expect  max.err error counts
3       952432  509772.4        0       952432
4       226667  127443.1        0       226667
5       43474   31860.8 0       43474
6       18566   7965.2  0       18566
7       4152    1991.3  0       4152
8       725     497.8   0       725
9       2218    124.5   0       234 1984
10      2876    31.1    1       61 2815
11      1095    7.8     1       21 1074
12      231     1.9     1       1 230
13      19      0.5     1       0 19
14      9       0.1     1       1 8
24      1       0.0     2       1
25      1       0.0     2       0 0 1
26      1       0.0     2       0 1
28      1       0.0     2       0 0 1
30      1       0.0     3       1
31      1       0.0     3       0 1
83      1       0.0     3       0 1
90      1       0.0     3       1
91      1       0.0     3       1
92      1       0.0     3       1
143     1       0.0     3       0 0 1
150     1       0.0     3       0 1

=== Second read: Adapter 2 ===

Sequence: AGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCGTATCTCGTATGCCGTCTTCTGCTTG; Type: regular 3
'; Length: 66; Trimmed: 1212532 times.

No. of allowed errors:
0-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-39 bp: 3; 40-49 bp: 4; 50-59 bp: 5; 60-66 bp: 6

Bases preceding removed adapters:
  A: 35.3%
  C: 22.3%
  G: 29.3%
  T: 13.1%
  none/other: 0.0%

Overview of removed sequences
length  count   expect  max.err error counts
3       823597  509772.4        0       823597
4       205320  127443.1        0       205320
5       63076   31860.8 0       63076
6       27599   7965.2  0       27599
7       24078   1991.3  0       24078
```



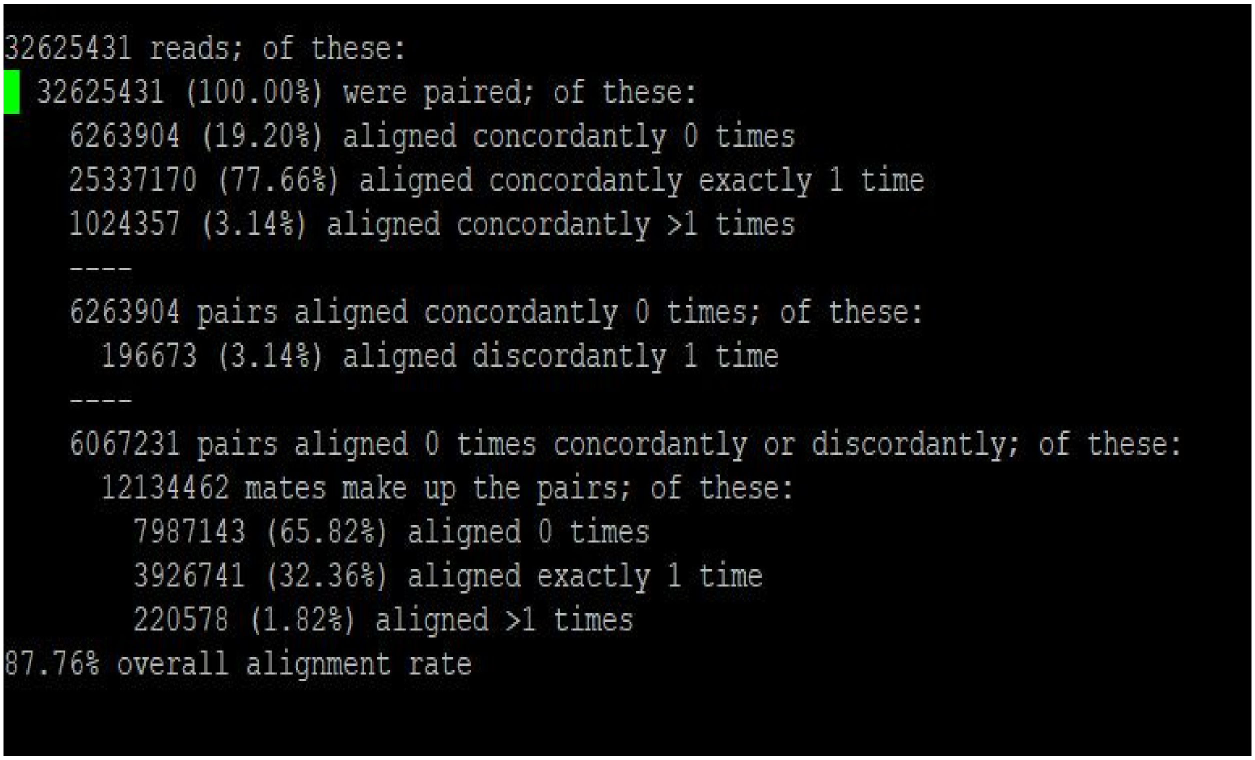
```
Overview of removed sequences
length  count   expect  max.err error counts
3       823597  509772.4        0       823597
4       205320  127443.1        0       205320
5       63076   31860.8 0       63076
6       27599   7965.2  0       27599
7       24078   1991.3  0       24078
8       22693   497.8   0       22693
9       22749   124.5   0       21527 1222
10      22303   31.1    1       20506 1797
11      438     7.8     1       63 375
12      314     1.9     1       118 196
13      132     0.5     1       102 30
14      229     0.1     1       0 229
16      2       0.0     1       0 2
34      1       0.0     3       1
35      1       0.0     3       1
```

Supplementary Figure 2. STAR alignment output

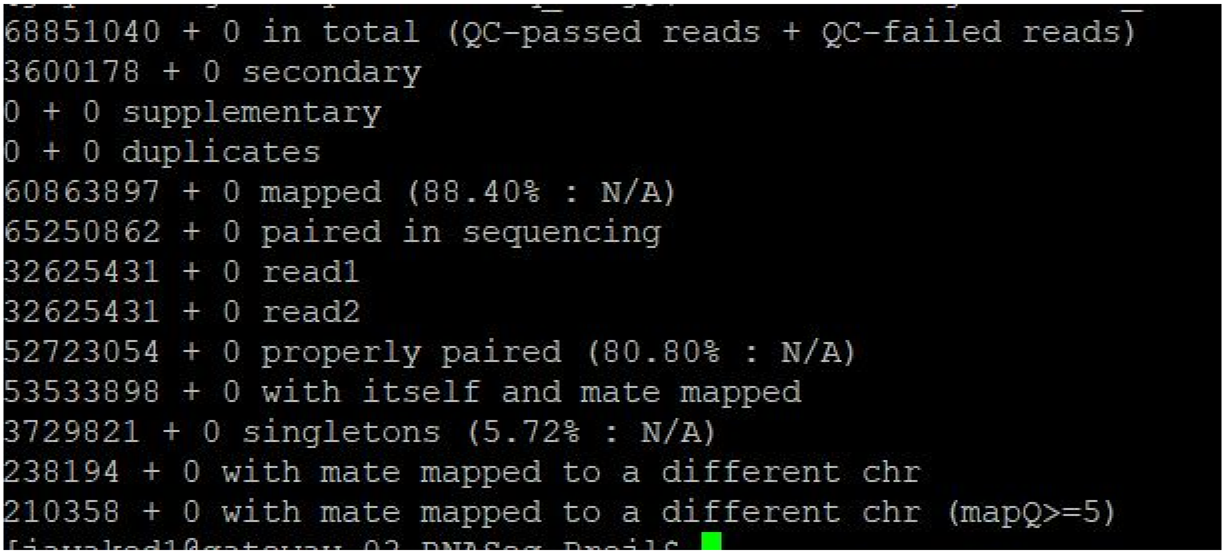


```
Mapping speed, Million of reads per hour |       86.81

                   Number of input reads |       32625431
               Average input read length |       300
                             UNIQUE READS:
            Uniquely mapped reads number |       29581094
                 Uniquely mapped reads % |       90.67%
                   Average mapped length |       296.06
                Number of splices: Total |       27760273
     Number of splices: Annotated (sjdb) |       20676879
                Number of splices: GT/AG |       27336439
                Number of splices: GC/AG |       263931
                Number of splices: AT/AC |       12771
        Number of splices: Non-canonical |       147132
               Mismatch rate per base, % |       0.98%
                  Deletion rate per base |       0.06%
                 Deletion average length |       2.93
                 Insertion rate per base |       0.05%
                Insertion average length |       2.82
                      MULTI-MAPPING READS:
 Number of reads mapped to multiple loci |       1615281
      % of reads mapped to multiple loci |       4.95%
 Number of reads mapped to too many loci |       12856
      % of reads mapped to too many loci |       0.04%
                           UNMAPPED READS:
% of reads unmapped: too many mismatches |       0.00%
          % of reads unmapped: too short |       4.31%
              % of reads unmapped: other |       0.03%
                           CHIMERIC READS:
                Number of chimeric reads |       0
                     % of chimeric reads |       0.00%
```

Supplementary figure 3- HISAT2 output

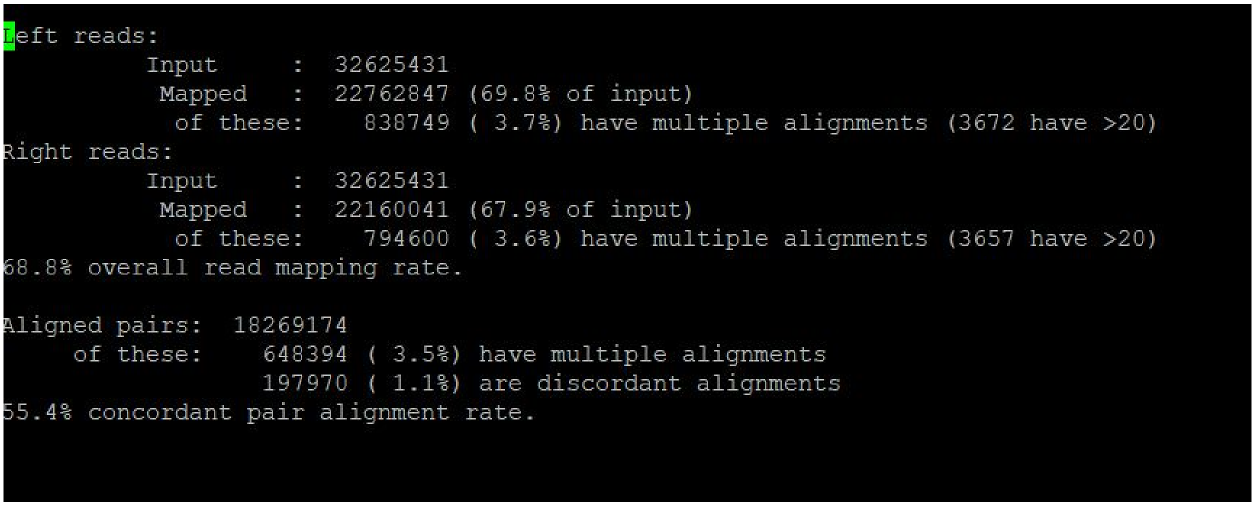


```
32625431 reads; of these:
  32625431 (100.00%) were paired; of these:
    6263904 (19.20%) aligned concordantly 0 times
    25337170 (77.66%) aligned concordantly exactly 1 time
    1024357 (3.14%) aligned concordantly >1 times
    ----
    6263904 pairs aligned concordantly 0 times; of these:
      196673 (3.14%) aligned discordantly 1 time
    ----
    6067231 pairs aligned 0 times concordantly or discordantly; of these:
      12134462 mates make up the pairs; of these:
        7987143 (65.82%) aligned 0 times
        3926741 (32.36%) aligned exactly 1 time
        220578 (1.82%) aligned >1 times
87.76% overall alignment rate
```



```
68851040 + 0 in total (QC-passed reads + QC-failed reads)
3600178 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
60863897 + 0 mapped (88.40% : N/A)
65250862 + 0 paired in sequencing
32625431 + 0 read1
32625431 + 0 read2
52723054 + 0 properly paired (80.80% : N/A)
53533898 + 0 with itself and mate mapped
3729821 + 0 singletons (5.72% : N/A)
238194 + 0 with mate mapped to a different chr
210358 + 0 with mate mapped to a different chr (mapQ>=5)
```

Supplementary figure 4- Tophat2 Summary output



```
Left reads:
          Input     :  32625431
           Mapped   :  22762847 (69.8% of input)
            of these:    838749 ( 3.7%) have multiple alignments (3672 have >20)
Right reads:
          Input     :  32625431
           Mapped   :  22160041 (67.9% of input)
            of these:    794600 ( 3.6%) have multiple alignments (3657 have >20)
68.8% overall read mapping rate.

Aligned pairs:  18269174
     of these:    648394 ( 3.5%) have multiple alignments
                  197970 ( 1.1%) are discordant alignments
55.4% concordant pair alignment rate.
```