SEXUAL DIMORPHIC GENE EXPRESSION IN MONOECIOUS AND DIOECIOUS HEMP (*Cannabis sativa*)

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Professional Studies

by

Ann Tate

December 2020

ABSTRACT

Sexual dimorphism is common among hemp (*Cannabis sativa* L.), with both monoecious and dioecious varieties arising. With the renewed interest in hemp breeding, it is important to increase our understanding of the genetics underlying these complex traits, as little is currently known. In this project, the gene expression of four different varieties of male and female dioecious and one male and female monoecious hemp variety was examined. The results of the gene expression analysis, provide candidate gene lists for future studies into monoecy in hemp, as well as a broad picture of gene expression across multiple cultivars.

BIOGRAPHICAL SKETCH

Ann Tate received her bachelor of science in biology and education from St. John Fisher College in Rochester, New York. Immediately after college she started working for the Cornell BRC genomics core as a laboratory technician. In 2018 she joined the Transcriptional Regulation.

ACKNOWLEDGMENTS

I would like to thank Larry Smart for taking me on as an MPS student and supporting me on this journey, and for always being available and present, even in the middle of a pandemic. I would also like to thank my boss Jen Grenier and lab mates Chrissy and Faraz, who were there to help and encourage every step of the way. I would particularly like to thank Jacob Toth, for his endless knowledge on all hemp related things, and for helping me to complete this project, I truly could not have done it without his help. Last but not least my dear friend Katie, who inspires me to be the best person I can be, and who taught me that continuing your education can happen at any age. Most importantly, I would like to than my husband, Ross, for always supporting me and believing in me, even when I did not believe in myself

TABLE OF CONTENTS

INTRODUCTION:

Hemp (*Cannabis sativa* L.) is a naturally dioecious crop species, with separate male (XY) and female (XX) plants, but monoecious phenotypes are common, especially among grain cultivars. This species is characterized by sexual dimorphism, particularly in floral ontogeny and plant height, and this dimorphism makes sex an important trait for hemp breeding. Monoecious cultivars are also of interest to farmers, as they tend to have more crop homogeneity making them easier to harvest with greater seed yields (Mandolino, 2004).

The chromosome set for hemp is composed of nine autosomes and one pair of sexual chromosomes, X and Y. Sex determination is believed to be based on an X:autosome dosage and not on a Y-active mechanism (Grant, 1994). The Y chromosome, like many male chromosomes, is strongly heterochromatic and rich in repetitive sequences. It is also larger than it's X counterpart, with male genomes being on average 2.7% larger than female genomes (Faux A. M., 2014). Monoecious cultivars are genetically XX, with a similar genome size to female hemp, and the absence of male-specific markers (Faux A. M., 2014).

With the legalization of hemp production in the United States through the Agricultural Improvement Act of 2018, there is renewed interest in hemp not only as an industrial crop, but also for medicinal use. In most cases, sex plays and key role in production for these emerging hemp markets and is thus an important consideration for breeders. For fiber, female plants are known to be more lignified than males, with males producing finer fibers (Liu, 2015). In grain production, female floral density and timing relative to male pollen release is critical for yield (Faux A. M., 2014). Most notably for CBD production female plants are preferred over males, because female inflorescences are known to accumulate far greater cannabinoid content than their male counterparts (Small, 2016).

Recent research into hemp genomics, including the release of several well annotated reference genomes has enabled more comprehensive genetic analysis. In this project we sought to create an RNA-Seq dataset of four different hemp cultivars: 'Anka', 'Logan', 'SC1', and 'Otto II', in order to better understand the genetics of sexual dimorphism and monoecy. We selected these cultivars to expand on previous work done in medicinal hemp (Braich, 2019) to include grain and fiber varieties from a broader range of cultivars. We also chose to analyze against both a male (XY) and female (XX) reference in order to look at differences between sexes. Our aim is to identify candidate genes associated with sex determination in dioecious cultivars and in regulating monoecy, as well as genes influencing sexual dimorphism in common or unique to each cultivar.

MATERIALS AND METHODS:

Four hemp cultivars 'Anka', 'Logan', 'Otto II', and 'SC1' were selected for RNA-Seq analysis and were grown in the summer of 2018 under the following conditions (Table 1).

Table 1:

Cultivar	**Seed Origin**	**Growth Conditions**
Anka	Uniseeds, Cobden, ON	Research Farm North
Logan	Improved feral population from NY (ID: GBAH-18-1024)	Greenhouse
Otto II	Winterfox Farm, Klamath Falls, OR	Gates West Farm
SC1	PreProcess, Inc, Ellisburg, NY	Crittendon North Farm

Dioecious male and female inflorescences from all cultivars, and monoecious inflorescences from 'Anka' were harvested at a mature flowering stage and flash frozen in liquid nitrogen. Male and female flowers were removed from the stem and a Genogrinder 2000 (Spex CertiPrep, Metuchen, NJ) was used to homogenize the tissue (Figure 1). For monoecious 'Anka' samples, male flowers were carefully separated from female flowers on the same stem before homogenization.

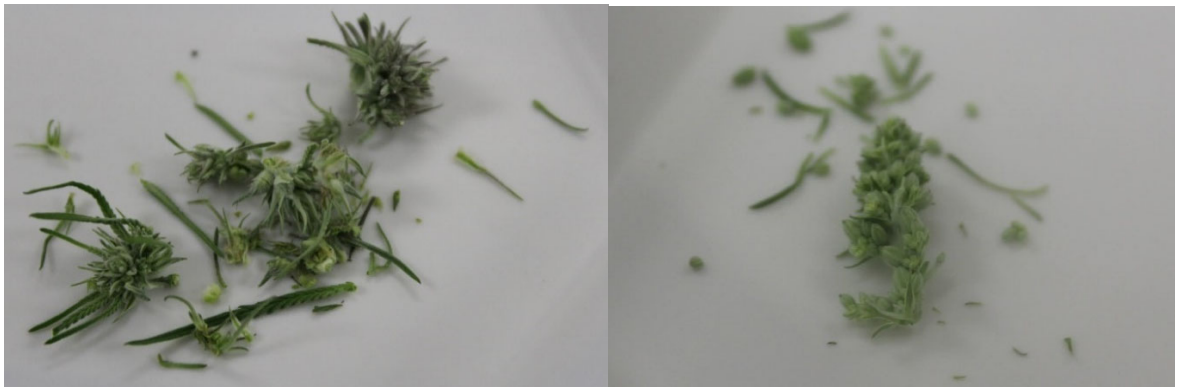


Figure 1: Female (left) and male (right) hemp flowers, prior to homogenization.

RNA was extracted using the Sigma Spectrum Total Plant RNA kit (Sigma-Aldrich, St. Louis, MO) with the modification at the binding step to capture small RNA molecules. DNA was removed with the RapidOUT DNA Removal kit (Thermo Fisher Scientific, Waltham. MA). RNA sample quality was confirmed by spectrophotometry (Nanodrop 8000, Thermo Fisher Scientific) to determine concentration and chemical purity (A260/230 and A260/280 ratios) and with a Fragment Analyzer (Agilent, Santa Clara, CA) to determine RNA integrity (Supplemental Figure 3). All samples had Nanodrop A260/230 and A260/280 ratios falling between 1.5 and 2. Samples with RNA Quality Numbers (RQN) lower than 5 on the Fragment Analyzer were removed in order to decrease 3' bias in the RNA library prep.

Poly A^+ RNA was isolated with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Ipswich, MA). TruSeq-barcoded RNA-Seq libraries were generated with the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs). Each library was quantified with a Qubit 4.0 (dsDNA HS kit; Thermo Fisher) and the size distribution was determined with a Fragment Analyzer prior to pooling (Supplemental Figure 4). Libraries were sequenced on a Illumina HiSeq 4000 instrument (San Diego, CA) at

a 2x150bp read length with a goal of 20 million raw reads per sample (read counts can be found in Appendix 1).

Reads were trimmed for low quality and adaptor sequences with TrimGalore v0.6.0 (Krueger, n.d.), a wrapper for cutadapt (Martin, 2011) and fastQC (Andrews, n.d.).

Parameters: -j 1 -e 0.1 --nextseq-trim=20 -O 1 -a AGATCGGAAGAGC --length 50 --fastqc

Unwanted reads were removed with STAR v 2.7.0e (Alexander Dobin, 2013) and the remaining reads were mapped to both a male (XY) and female (XX) genome, Jamaican Lion father (NCBI: JL_father) and *Cannabis sativa* (ensembl: cs10) respectively, using STAR v2.7.0e (Alexander Dobin, 2013).

Parameters: --outReadsUnmapped Fastx

Parameters: --outSAMstrandField intronMotif , --outFilterIntronMotifs RemoveNoncanonical , --outSAMtype BAM SortedByCoordinate, --quantMode GeneCounts

SARTools (Hugo Varet, 2016) and DESeq2 v1.26.0 (Love, 2014) were used to generate normalized counts, statistical analysis of differential gene expression, and PCA graphs.

Parameters: fitType parametric, cooksCutoff TRUE, independentFiltering TRUE, alpha 0.05, pAdjustMethod BH, typeTrans VST, locfunc median

Differential gene expression analysis was performed with six different comparisons; all males (including monoecious males) to all females (including monoecious females), male to female within the four cultivars, and monecious male ‘Anka’ to dioecious male ‘Anka’

(Appendix 2). Figures were made using ggplot2 in R (Wickham, 2016). Heatmapper was used to create the heatmaps using Pearson's distance measurement (Sasha Babicki, 2016)

RESULTS & DISCUSSION:

Reads aligned to CS10 with an average of 85% alignment (chromosome level assembly), and to JL_father (contig level assembly) with an average of 70.5% alignment (Supplemental figures 1 & 2, Appendix 1). An average of 30 million reads were obtained per sample. Gene Body coverage was examined to rule out any 3' bias (Figure 2).

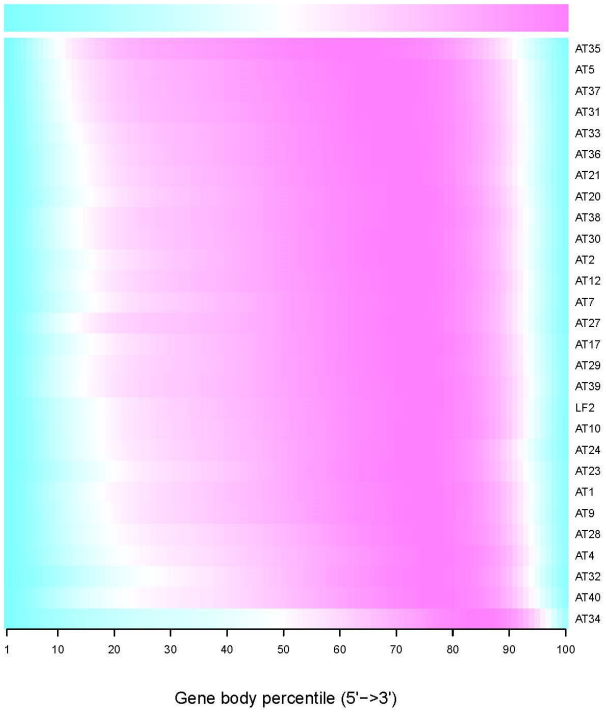


Figure 2: Gene Body coverage heatmap showing relative consistency among samples, with the exception of AT34-SC1 rep2. Other QC metrics for this sample show it in alignment with it's group members, indicating the 3' bias did not affect the results.

The principal components analysis (PCA) for both the CS10 and JL_father shows good clustering among replicates and appears to separate male and female samples on PC1

and cultivars on PC2 (Figure 3). Notably, in ‘Anka’, monoecious females clustered closely with dioecious females and monoecious males clustered more closely with dioecious males.

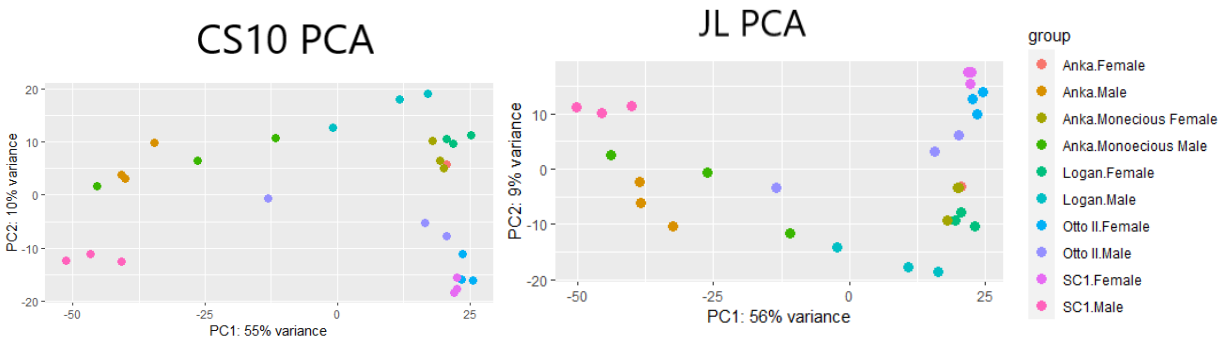


Figure 3: PCA’s of CS10 and JL analysis with PC1 separating out by sex and PC2 separating out cultivars. Close replicate grouping indicates similarity between replicates and good data quality.

Differential expression analysis across CS10 and JL yielded many differentially expressed genes across all comparisons (Appendix 2, Supplemental Figures 3&4).

There were more genes differentially expressed in females than males (Supplemental Figure 3, and Figures 4 & 5). The JL_father (XY) analysis and CS10 (XX) analysis yielded quite different results in terms of both the number of differentially expressed genes and the types of genes differentially expressed. This may be due to differences in the alignment, as reads aligned to CS10 at a higher rate than to JL, but could also be due to the addition of the Y chromosome in the JL_father alignment.

Both analysis showed distinct differences in the replicates of ‘Logan’ and ‘Otto II’ male samples, which also is visible in the PCA (Figures 3-5). In looking at the differentially expressed genes across these replicates, there were several genes associated with development and defense in the list. This could mean that flowers were selected at differential stages of growth or that the plant was under stress from a pathogen when it was harvested. More replicates would be needed to determine what factors were underlying these differences.

In the JL_father alignment there is a set of genes that appear differentially expressed in the monecious male as compared to the dioecious (XY) male (Figure 4, Supplemental Figure 7). In this gene set, the expression of the monoecious male is more in alignment with female gene expression than male. This could suggest that these genes fall on the Y chromosome, as they are expressed at a much higher rate in the dioecious males than in both the monoecious males and females, and monoecious males are known to be XX (Faux A. M., 2014). These differences are missing in the CS10 (XX) male vs. female differential gene expression, likely due to the lack of a Y chromosome in the CS10 analysis (Figure 5).

In conclusion, differential expression analysis across CS10 and JL_father yielded many differentially expressed genes across all comparisons (Appendix 2, Supplemental Figures 3&4). With a focus on monoecy, there is a candidate gene list for the Y chromosome of JL_father as well as a list of candidate genes for monoecy. There is a wealth of knowledge still to be gained from this dataset, but this is a good initial start to the analysis .

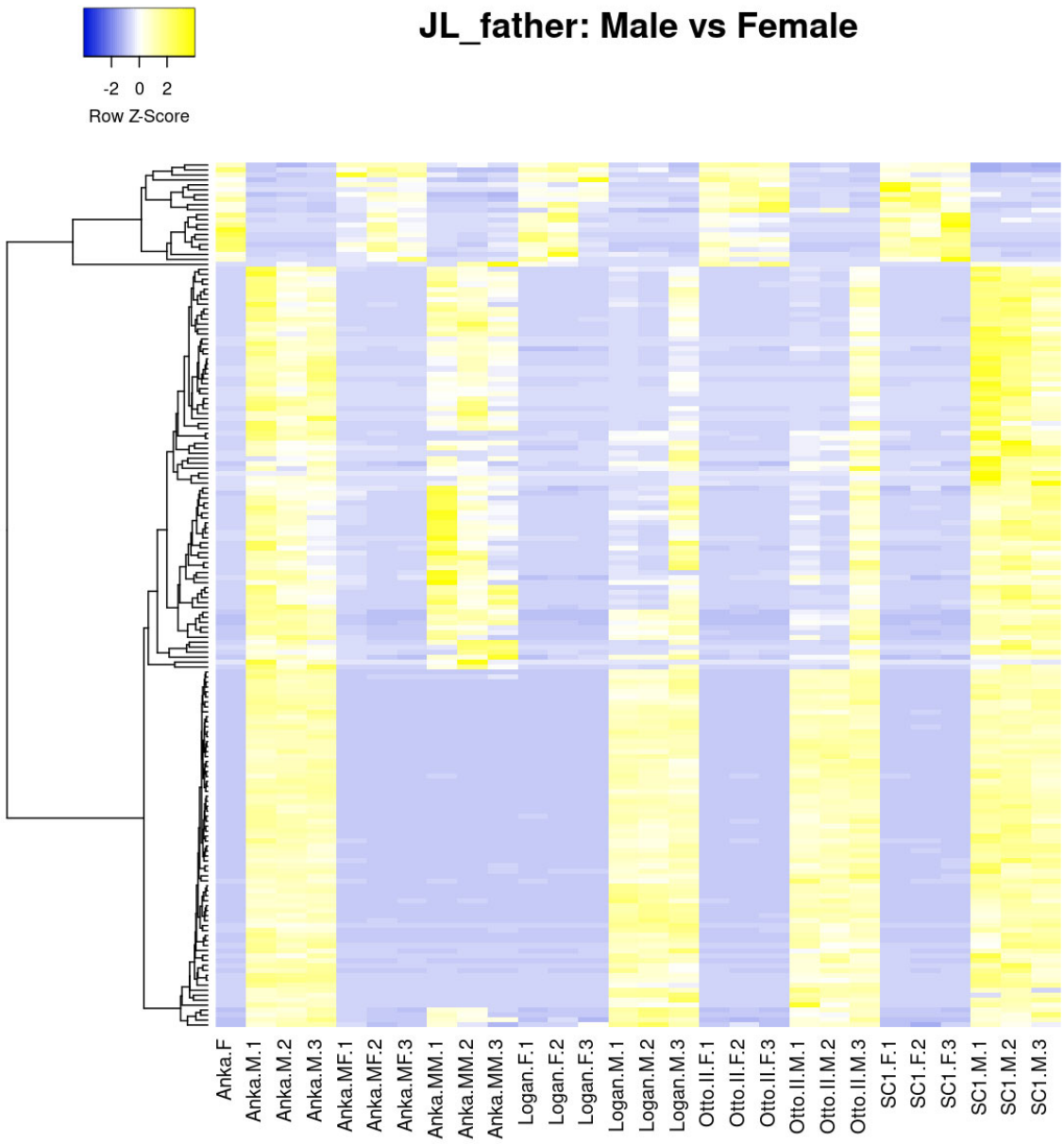


Figure 4: Heatmap of 176 genes differentially expressed in all male and female cultivars, where monoecious males (MM) were grouped with dioecious males (M) and monoecious females (MF) were grouped with dioecious females (F), when aligned to JL_father.

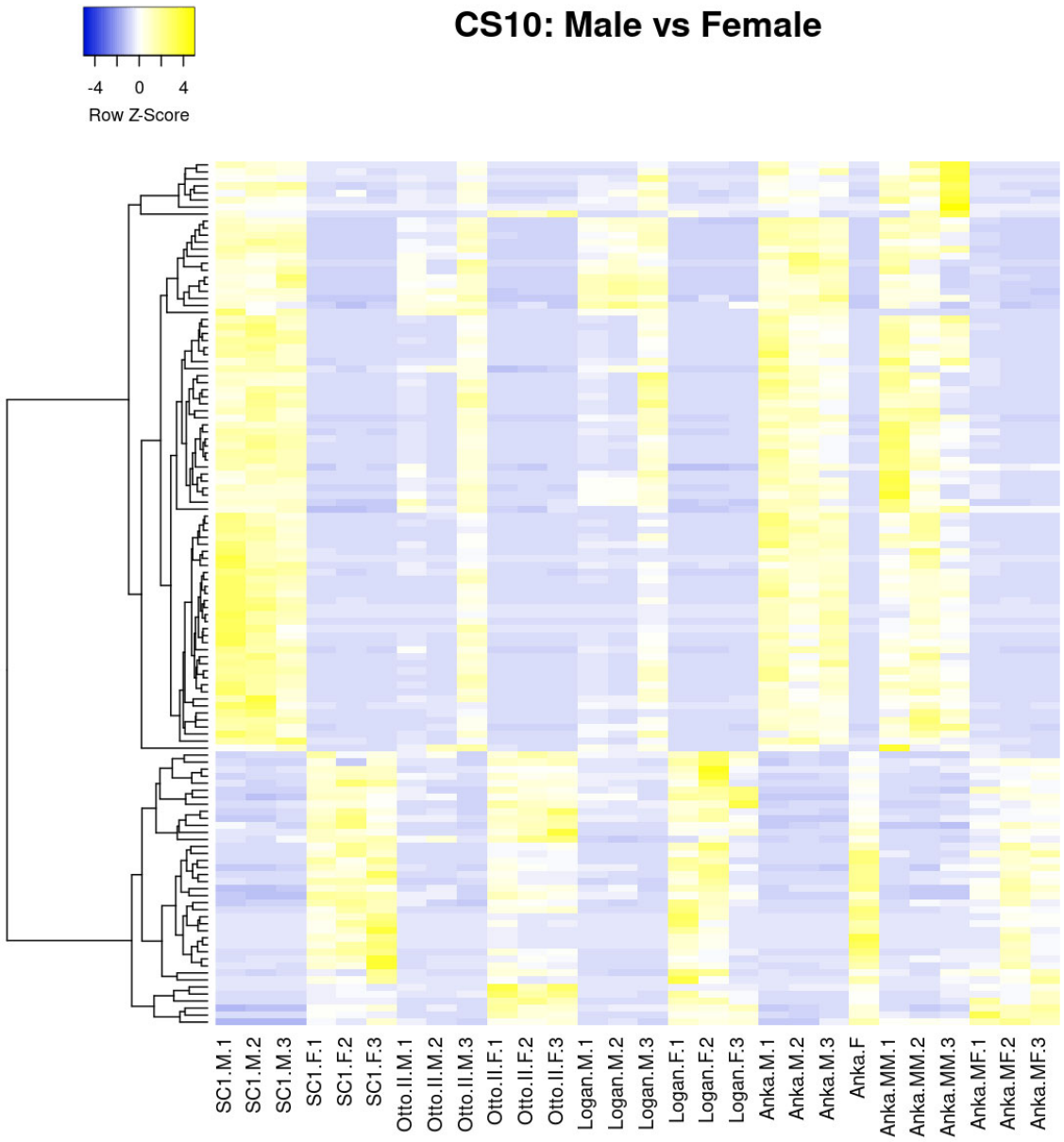


Figure 5: Heatmap of 123 genes differentially expressed in all male and female cultivars, where monoecious males (MM) were grouped with dioecious males (M) and monoecious females (MF) were grouped with dioecious females (F), when aligned to CS10.

REFERENCES:

Alexander Dobin, C. A. (2013, January). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 15-21. Retrieved from https://doi.org/10.1093/bioinformatics/bts635

Andrews, S. (n.d.). *fastqc*. Retrieved from Babraham Bioinformatics: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Braich, S. B. (2019). Generation of a Comprehensive Transcriptome Atlas and Transcriptome Dynamics in Medicinal Cannabis. *Scientific Reports*. doi:https://doi.org/10.1038/s41598-019-53023-6

Faux, A. M. (2014). Modelling approach for the quantitative variation of sex expressions in monoecious hemp (Cannabis sativa L.). *Plant Breeding*, 782-787.

Faux, A. M. (2014). Sex chromosomes and quantitative sex expression in monoecious hemp (Cannabis sativa L.). *Euphytica*, 183-197.

Grant, S. A. (1994). Genetics of sex determination in flowering plants. *genesis, 15*(3), 214-230.

Hugo Varet, L. B.-G.-Y.-A. (2016, June 9). SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLOS*. Retrieved from http://dx.doi.org/10.1371/journal.pone.0157022

Krueger, F. (n.d.). *TrimGalore*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Liu, M. F. (2015). Effect of harvest time and field retting duration on the chemical composition, morphology and mechanical properties of hemp fibers. *Industrial Crops and Products*, 29-39.

Love, M. H. (2014, December 5). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*. Retrieved from https://doi.org/10.1186/s13059-014-0550-8

Mandolino, G. C. (2004). Potential of marker-assisted selection in hemp genetic improvement. *Euphytica*, 107-120.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 10-12.

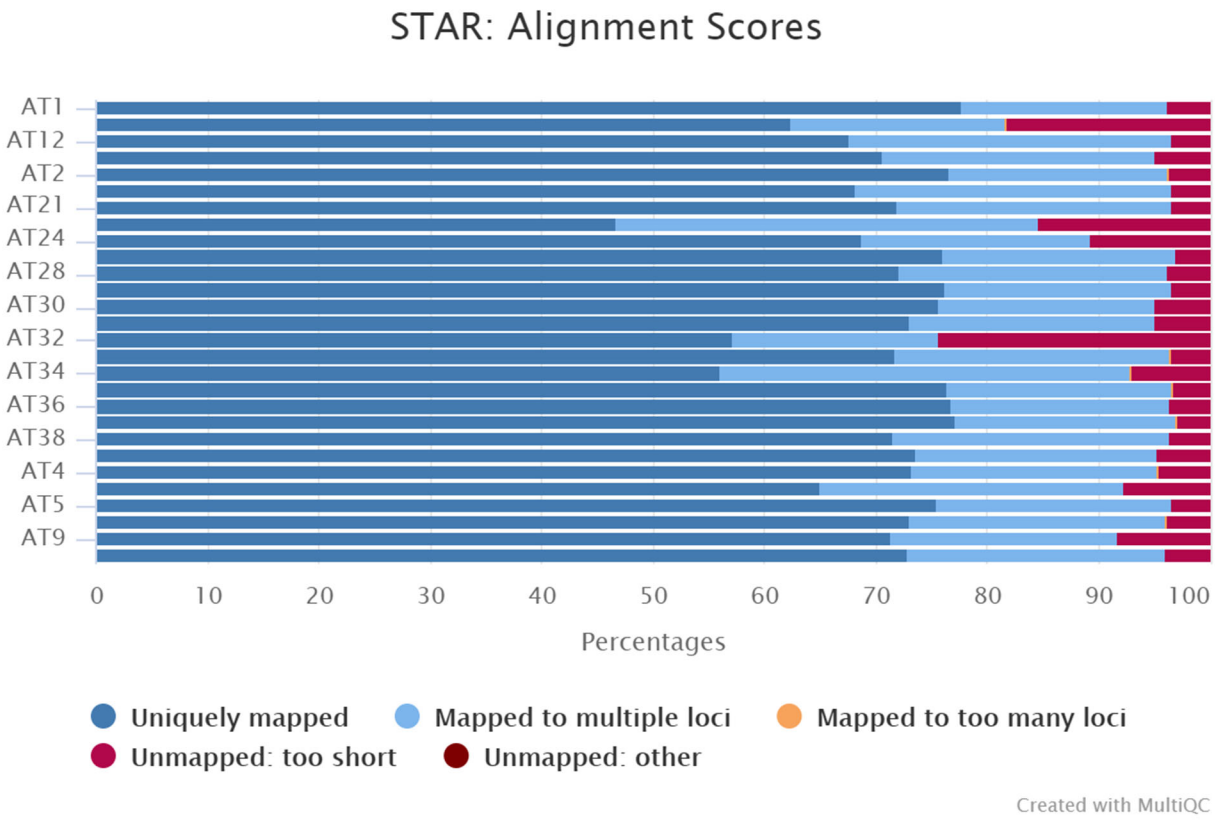
Salentijn, E. M. (2015, June). New developments in fiber hemp (Cannabis sativa L.) breeding. *Industrial Crops and Products*, 32-41.

Sasha Babicki, D. A. (2016). Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* doi:doi:10.1093/nar/gkw419

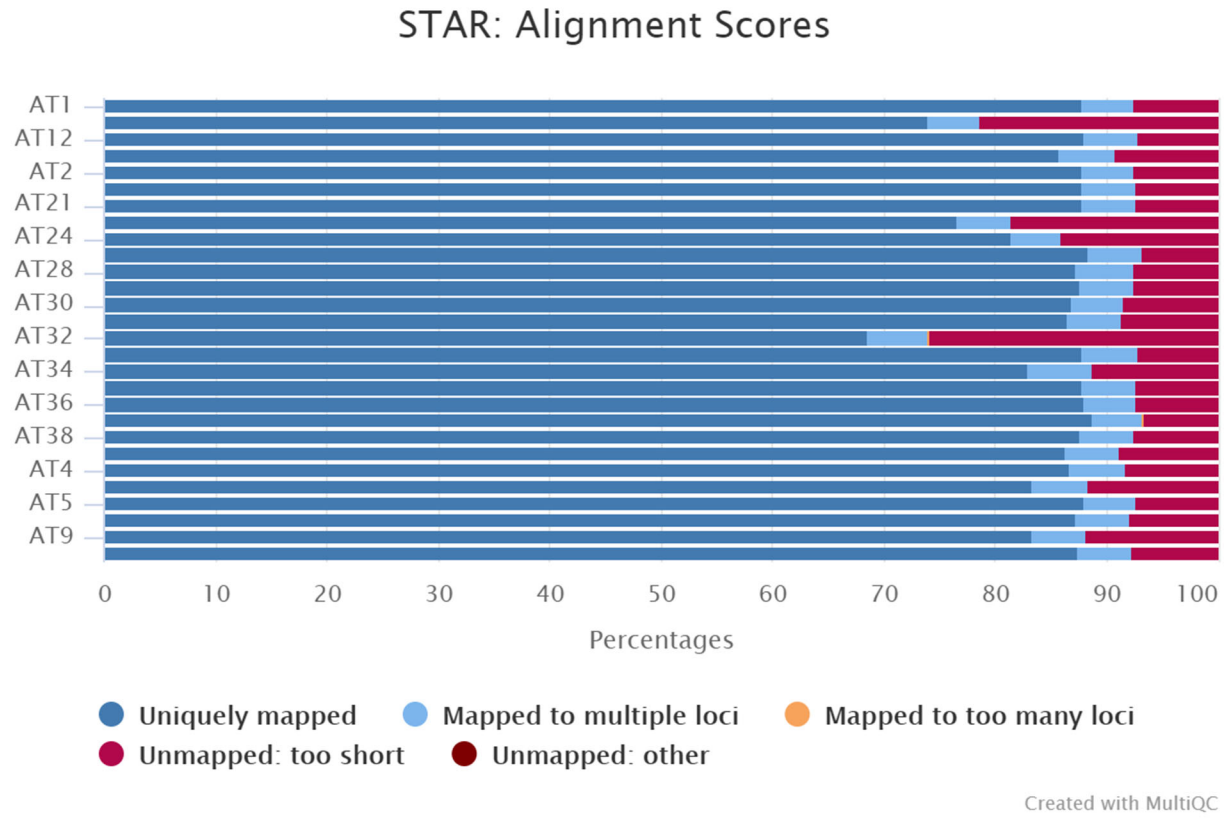
Small, E. N. (2016). Expansion of female sex organs in response to prolonged virginity in Cannabis sativa (marijuana). *Genetic Resources and Crop Evolution*, 339-348.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved from https://ggplot2.tidyverse.org

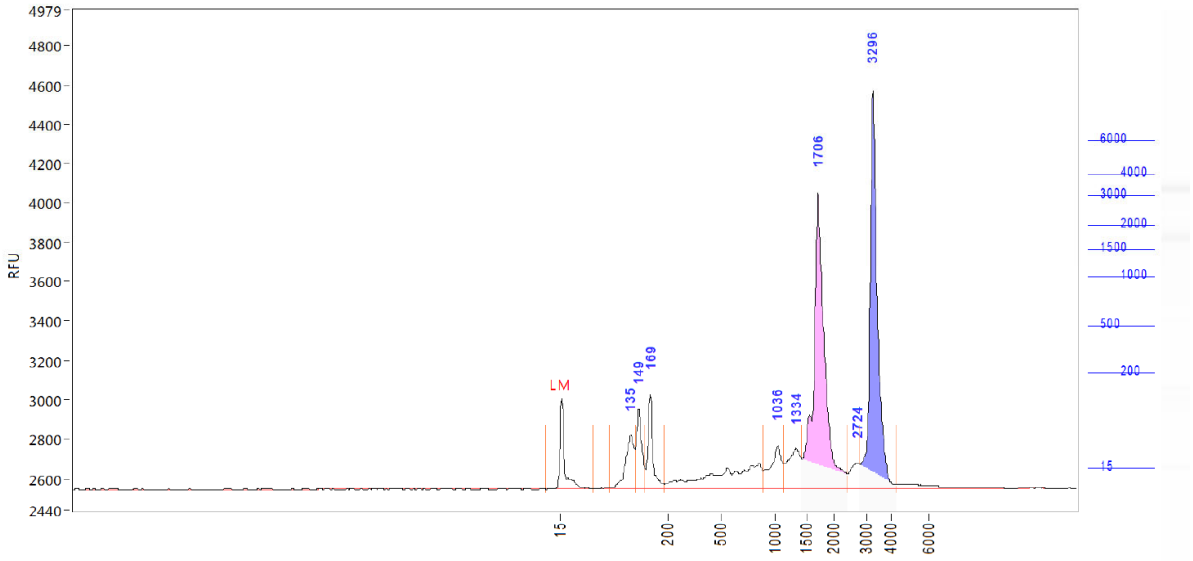
SUPPLEMENTAL FIGURES:



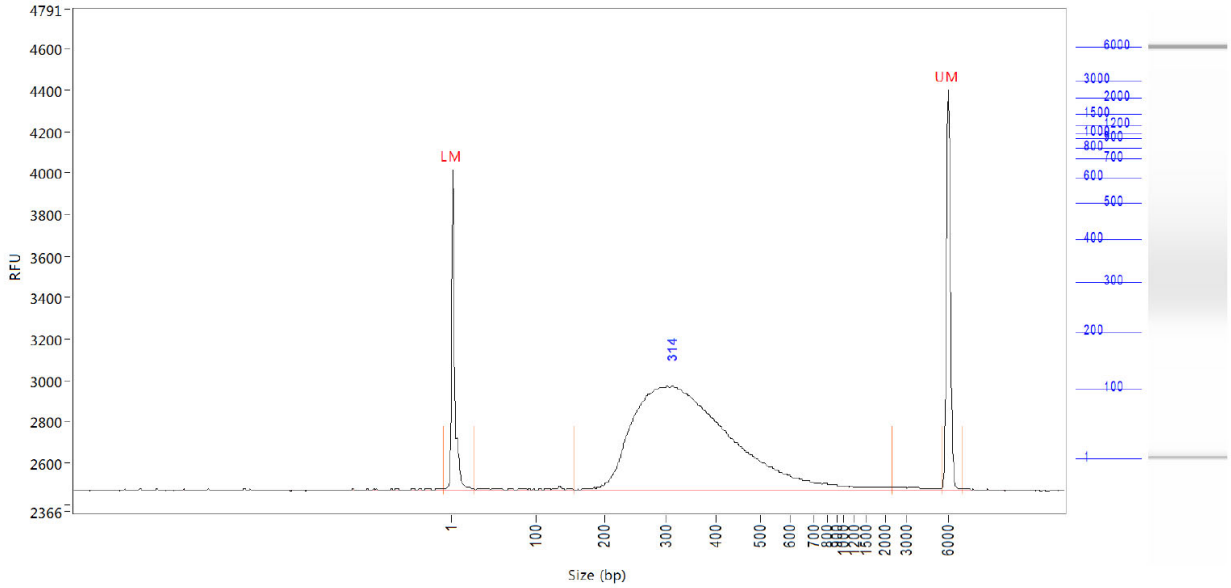
Supplemental Figure 1: STAR alignment of reads with JL, showing an average of 70% alignment across all cultivars.



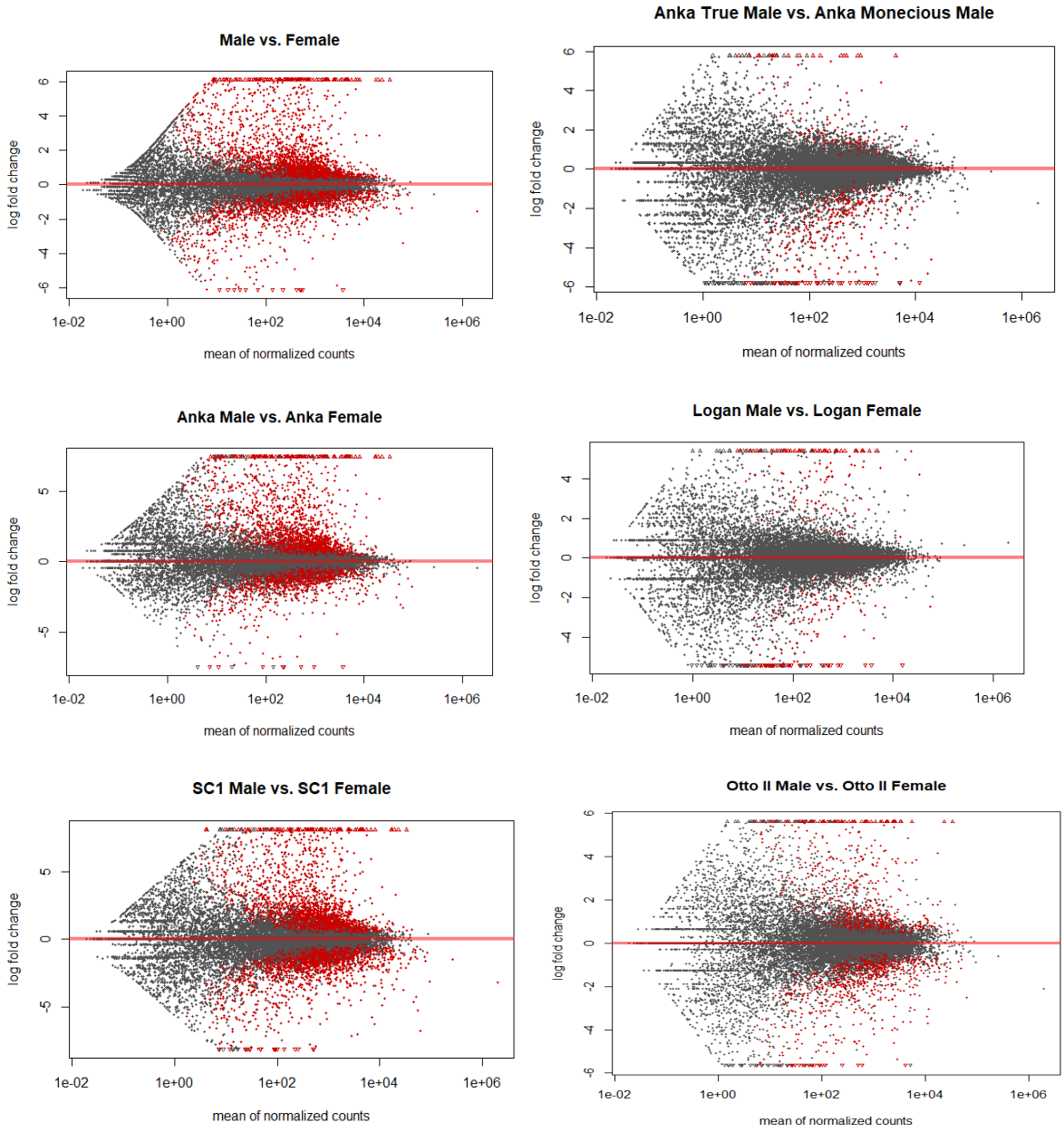
Supplemental Figure 2: STAR alignment with CS10, showing an average of 85% alignment across all cultivars.



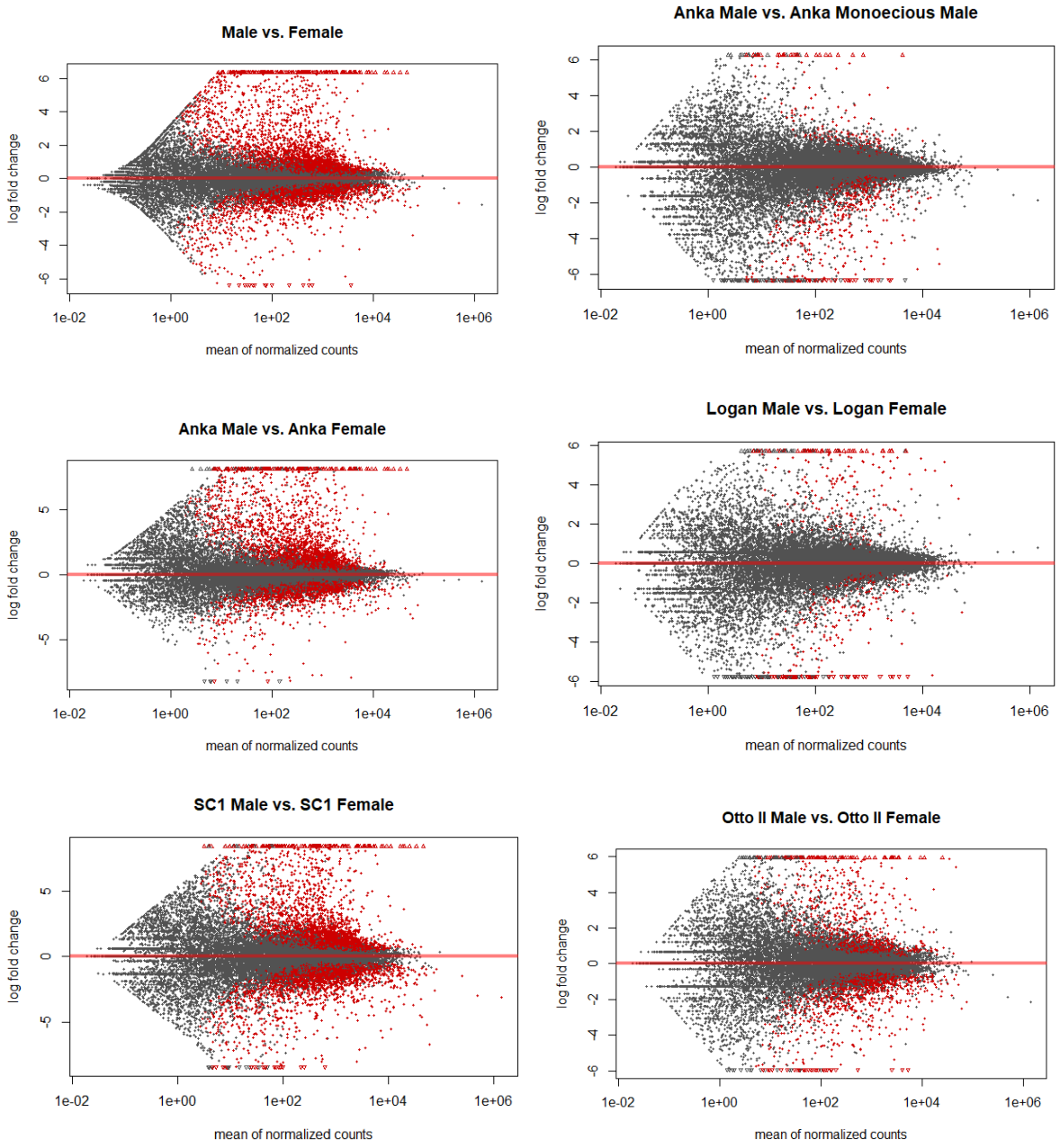
Supplemental Figure 3: Example RNA QC trace image from the Fragment Analyzer with an RNA Quality Number (RQN) of 8.1, showing intact RNA and no DNA contamination



Supplemental Figure 4: Example Fragment Analyzer trace of a representative RNA-Seq library. Absence of peak at ~150 bp indicates adapter dimer was removed, and absence of a peak ~30 bp indicates there is no leftover free primer demonstrating a clean library.



Supplemental Figure 5: Dispersion graphs for all comparisons, with red dots representing genes that are statistically differentially expressed.



Supplemental Figure 6: Dispersion graphs for all comparisons, with red dots representing genes that are statistically differentially expressed.

Supplemental Figure 7: List of genes differentially expressed in monoecious males as compared to dioecious males when analyzed with JL_father.

G4B88_024475	G4B88_005174	G4B88_001376	G4B88_011067
G4B88_008825	G4B88_007368	G4B88_003184	G4B88_015924
G4B88_013154	G4B88_003183	G4B88_001065	G4B88_031305
G4B88_031307	G4B88_008538	G4B88_016080	G4B88_009424
G4B88_000958	G4B88_012804	G4B88_018807	G4B88_011069
G4B88_029624	G4B88_020320	G4B88_029244	G4B88_017182
G4B88_009958	G4B88_005810	G4B88_001374	G4B88_001067
G4B88_003881	G4B88_021008	G4B88_020943	G4B88_001069
G4B88_018836	G4B88_029622	G4B88_009418	G4B88_015932
G4B88_001068	G4B88_000695	G4B88_016070	G4B88_015928
G4B88_018202	G4B88_000959	G4B88_012806	G4B88_009398
G4B88_030790	G4B88_001377	G4B88_003394	G4B88_009425
G4B88_027481	G4B88_028211	G4B88_015927	G4B88_000953
G4B88_004997	G4B88_021005	G4B88_016075	G4B88_030273
G4B88_009871	G4B88_001370	G4B88_018069	G4B88_009461
G4B88_009429	G4B88_009400	G4B88_010141	G4B88_008546
G4B88_027480	G4B88_000960	G4B88_030792	G4B88_025235