**The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:**

| | |
|---|---|
| **Document Title:** | **Genomic Tools to Reduce Error in PMI Estimates Derived from Entomological Evidence** |
| **Author(s):** | **Aaron M. Tarone, Christine Picard, Sing-Hoi Sze** |
| **Document No.:** | **250086** |
| **Date Received:** | **July 2016** |
| **Award Number:** | **2012-DN-BX-K024** |

**This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this federally funded grant report available electronically.**

**Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.**

| | |
|---|---|
| Federal Agency & Organization Element to Which Report is Submitted: | NIJ FY 12 Research and Development on Instrumental Analysis for Forensic Science Applications |
| Federal Grant or Other Identifying Number Assigned by Agency: | 2012-DN-BX-K024 |
| Project Title: | Genomic tools to reduce error in PMI estimates derived from entomological evidence |
| PD/PI Name, Title and Contact Information: | Aaron M. Tarone<br>2475 TAMU<br>Department of Entomology<br>Texas A&M University<br>College Station, TX 77843-2475 |
| Name of Submitting Official, Title, Contact information if other than PD/PI: | NA |
| DUNS and EIN Numbers: | DUNS: 847205713<br>EIN: 74-6000541 |
| Recipient Organization (Name & Address): | Texas A&M Research Foundation<br>400 Harvey Mitchell Parkway, Suite 100<br>College Station, TX 77845 |
| Recipient Identifying Number or Account Number: | 504561 |
| Project/Grant Period (Start Date, End Date): | 1/01/2012 through 12/31/2015 |
| Reporting Period End Date: | 12/31/2015 |
| Report: | Draft Final Summary Overview |
| Signature of Submitting Official: | |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

## Introduction and Rationale

Forensic entomologists estimate the ages of the immature insects like blow flies, which can provide useful information regarding timelines in death investigations (Byrd and Castner 2010). While this information is useful, it is also clear that there are ways to improve both the accuracy and precision of estimates with blow fly age through the use of genetic approaches (Tomberlin et al. 2011a, Tomberlin et al. 2011b). Estimates typically rely on life history traits of carrion flies, which are quantitative traits that are known to vary due to genetic and environmental factors in insects. Recently, experiments have demonstrated conspecific genetic variation in blow fly development (Gallagher et al. 2010, Tarone et al. 2011, Owings et al. 2014); confirming that carrion flies are likely not exceptions to numerous observations of genetic variation in insects (Mousseau and Roff 1987, Roff and Mousseau 1987, Mousseau and Dingle 1991, Blanckenhorn 1997, 1998). Unfortunately, little is known about the consequences of genetic variation in blow fly traits of forensic relevance beyond its existence. In addition, estimates of fly age can vary considerably in their precision. For instance, pupation typically consists of approximately the last half of development (Tarone and Foran 2008). Unless other information is used by investigators, predictions of age with this and similar stages produce imprecise estimates of age - even when accurate. The research goals of the proposal were to obtain quantitative and functional genetic information for the blow fly *Cochliomyia macellaria* Fabricius (Diptera: Calliphoridae) – a common forensic indicator species. Ultimately this information can be used to develop both short and long term strategies for using genetic tools to account for uncertainty (with respect to both accuracy and precision) in forensic estimates of blow fly age.

**Aim 1: Genetics of development time variation.** Our principle approach for addressing concerns of inaccuracy due to genetic variation in blow fly development was to conduct a selection experiment on development time. This allowed us to observe the full distribution of development times for starting populations and to observe the change in means and variances in development time over tens of generations of selection. The resulting material from this selection experiment was then also sequenced using next-generation sequencing technology to simultaneously develop reference genomes from selected lines and to track allele frequency changes over time from each selection group.

## Selection Experiment

The selection experiment was conducted by collecting *C. macellaria* from different locations in Texas. In the first replicate run of the experiment, three populations were founded from >100 flies caught in College Station, Snook, and Longview, Texas. These populations were brought into the lab and ~1,200 of their offspring were reared at 25ºC. The



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

development time of each member of the cohort was recorded by observing eclosion three times a day. From this original baseline cohort, two separate lines were founded; one from the 200 fastest developing of the 1,200 and one from the slowest developing 200 adults. Between generations 23 and 28 in the slow selection regime (generations began to stagger over the duration of the experiment) selection was relaxed (by decreasing temporal resolution of sampling) and development times were not recorded due to logistical constraints of strain maintenance.

In order to evaluate the stability of results from the first run of the experiment, an additional replicate was founded in a subsequent year from individuals in College Station, Longview, and San Marcos (from 100, 100, and 56 founding individuals, respectively). The experiment was generally run in the same manner as the previous replicate; however in this instance a control population was also maintained.

Summaries of developmental variation in the experiment can be found in Figures 1-5 and Tables 1-2. At current count (experiments are ongoing) there have been 191,119 development times measured for the first replicate of selection and 199,287 from the second replicate. In both replicates development times ranged from 221-329 hours in the first generation, while the total range in development after 29-43 generations of selection (depending on the selection regime, replicate, and experimental population) was 148-504 hours. After 23+ generations of selection the fast selection group developed significantly faster than the slow selection group (Tukey's HSD, $p = 0.02$ in replicate 1 and $p < 0.000009$ in replicate 2). In the second replicate, selected lines developed differently from their control populations (Tukey's HSD, $p < 0.0005$), while control groups did not diverge from the original founders (Tukey's HSD, $p = 0.92$). Heritability scores were also calculated from the selection groups using the Breeder's Equation (Falconer 1989, Conner and Hartl 2004). Development times exhibited heritability scores of 0.08 - 0.27 for fast selection regimes and 0.07 - 0.24 for slow selection regimes.

**Thermal Responses of Selected Strains**

While our experiment was done at one temperature, flies in casework are wild and experience a range of temperature exposures. Accordingly, we have begun investigating the nature of thermal plasticity in the selected lines. This was done by raising egg masses of the same age from the selection groups, from three successive generations (after 20 generations of selection at 25ºC); in 20ºC, 25ºC, and 30ºC at the same time. Specifically, pupal mass, development time (from when eggs were laid until adults eclosed), and immature viability were measured in all temperatures as described in Owings et al. (2014). Results from this experiment are found in Figure 6. Most importantly, thermal



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

responses of the two selection regimes clearly differ and selection on development time also impacted body size (another forensically important trait) as well as larval viability.

## Whole Genome Sequencing

All DNA sequencing, genome assembly and comparative genomics, and candidate gene approaches to selection were done at the Picard Lab at Indiana University Purdue University Indianapolis (IUPUI) campus in Indianapolis. Preserved flies (N = 50) from each of the selected lines (Longview, College Station, and Snook) from the first replicate of selection at generation 26, as well as the original starting populations (N = 50), were subjected to DNA extractions (Qiagen DNeasy Blood and Tissue kit) and pooled for whole genome sequencing. Illumina HiSeq produced approximately 20-22X coverage for each selection group, which could be assembled into draft genomes. Summaries of genomic analyses can be found in Figures 7-11 and Tables 3-9.

## Whole Genome Assembly

A total of 12 draft assemblies were generated from the above data (Table 3). Data are only reported for Longview and College Station, as the results for Snook require further analyses to confirm the quality of the data before results can be reported. Each assembly was originally done using only the short read data using the commercially available assembler CLC Genomics Workbench v 8.03 (Qiagen). Each assembly was then evaluated based on common assembly metrics (Tables 4, 7- 9).

## Candidate Gene Approach and Validation

The goal of this project was to investigate the molecular mechanisms that govern development rate in *C. macellaria* by utilizing a candidate gene approach. Specifically, to identify homologs of *Drosophila melanogaster* genes known to impact development, and look for differences in their gene structures depending on their selection regime. These structural differences can be leveraged to provide a simple PCR-based tool for the identification of a molecular marker that is correlated with development time.

There were 47 candidate genes identified in *Drosophila* (using the gene ontology terms developmental growth under biological processes in www.flybase.org). Sequences from these genes were used to identify homologs in the *C. macellaria* draft genomes. These homologs were discovered in our two combined draft genomes (Cmbd-F and Cmbd-S) using a BLAST algorithm (www.blast.ncbi.nih.gov/Blast). Of these 47 genes, our genomes had homology to 33. Each candidate was then put through an *ab initio* gene prediction program (Augustus, www.bioinf.uni-greifwald.de/augustus) to



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

generate the most probable gene structure (see our pipeline in Figure 7, an example in Figure 8). Of these 33 candidate genes, 14 had complete gene predictions in both the fast and slow genomes. Alternatively, though some did not have gene predictions for both the fast and slow genomes, they had interesting candidate molecular markers that we followed up on. Following predictions, sequence alignments were completed (MUSCLE, www.ebi.ac.uk/Tools/msa/muscle) and eight candidate genes showed some level of polymorphisms (whether in introns, SNPs, or as InDels) (Figure 9). All predicted protein sequences were compared to Dipteran sequences for validation of gene structure using BLAST.

Several potentially informative polymorphisms were observed. In the coding sequence of the gene *happyhour*, a SNP results in a non-synonymous mutation, however, the amino acid change is a leucine-isoleucine change, which does not likely result in much difference in terms of function (however, the 3D structure of the protein should be generated to verify the result). A summary of these results are in Table 5. Of the 8 candidate genes for which we empirically predicted polymorphisms between the fast and slow genomes, 4 were selected for primer design: *Bitesize* (*btsz*), *Insulin-like Receptor* (*InR*), *Translationally controlled tumor protein* (*Tctp*), and *Target of rapamycin* (*Tor*). The primers were designed to either show a simple gel-based assay for differences in product length (i.e. InDels), or amplifications for downstream Sanger sequencing. A summary of the results are shown in Table 6. For the *InR* locus (InR_2), the expected difference between the slow (lane 1) and fast (lane 2) selection was 11 bp, which can be seen on the gel image below (Figure 10). Further validation can help to determine the role of this and similar polymorphisms on development time variation.

**Comparative Genomics Approach and Validation**

For our comparative genomics approach, we employed a strategy in which all nine libraries (reads) were mapped to the SLOW combined (Cmbd-S) genome. Once mapped, we could then extract variants (SNPs, MNVs, and InDels) and perform iterative analyses to extract variants common to all three geographic locations and linked to the selection regime. Mapping of the baseline populations was done to estimate the starting allele frequencies.

From the list of differentially fixed (between the fast and slow genomes) alleles (Figure 11, Tables 8-9), three loci were further selected for validation sequencing (PCR amplification of the individual loci across many generations of selection and Sanger sequencing was completed to validate our empirically derived frequencies). For example, one locus in the original baseline population (Longview) had an initial starting allele frequency of 0.25 (T) and 0.75 (C). After five generations in the FAST selected line, the frequency of the C allele was fixed (1.0), and after 10 generations in the SLOW



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

selected line, the SNP (T) had fixed (1.0). The same SNP in another population (College Station), with the same starting allele frequency, fixed in the SLOW population after 26 generations, meanwhile a new allele arose in the FAST population (however, the T allele was lost after only 10 generations). Therefore, with this preliminary data, it appears that the loss of the T allele may be correlated to a decrease in development rate, but this work will need further work to validate in additional independent populations.

## Future Directions

In order to ensure robust genomic analyses, it is necessary to assemble the genome as best as possible given resources available. Accordingly, long read sequence data (Pacific Biosciences) was generated at the Icahn Institute for Genomics and Multiscale Biology using a single male individual. This long read sequence data produced ~173,000 reads, averaging 3 Kb in length (approximately 2X coverage). We are currently working a with colleague to do hybrid genome assembly using deep short read sequences along with shallow long read data. A second draft genome will be generated for each of the 12 libraries listed above. Once these genomes are assembled, similar approaches to marker discovery as described above will be determined, including transcriptomic sequences generated from the RNAseq experiments to better describe and annotate the genome(s).

## Forensic Implications and Discussion

Recent publications in forensic entomology have opened a genetic "Pandora's Box". There are multiple experiments suggesting a genetic component to variation in forensic indicator traits of blow flies, but there is not much information regarding the impact of this variation on error in forensic entomology. Gallagher et al. (2010) did show that use of an incorrectly assumed development data set could lead to as much as ~14% error in insect age estimates, but that study was on three regional strains and thus was not an exhaustive consideration of genetic variation in blow flies. Clearly there is further need to account for impact of genetics on accuracy of results. This project has advanced our understanding of the role of genetics in uncertainty in forensic estimates of insect age. First, we have been able to show that from standing natural genetic variation, there is a genetic potential to drive average development time differences of approximately six days at 25ºC. This value represents a maximum error due to genetics. However, the genetic combinations produced in the selection experiment are not found in nature. This comparison is analogous to predicting phenotypic differences among wolves by observing the ability to produce Chihuahuas and Great Danes. This is an extremely conservative value.



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

An additional piece of information from this experiment was the use of the Breeder's Equation to estimate heritability. In a rudimentary sense, this calculation represents the degree to which offspring look like their parents. These results suggest that 7-27% of variation in *C. macellaria* development time is due to additive genetic variation, which appeared to be a stable result across selection regimes and replicate selection trials. This value is consistent with the heritability values for other insect development times (Mousseau and Roff 1987) and is consistent with other reports of error in forensic entomology (VanLaerhoven 2008, Gallagher et al. 2010). Accordingly, heritability estimates in blow flies may provide an empirical means by which investigators can express expectations of genetic sources of uncertainty in forensic entomology. It should be noted though, that heritability estimates are notoriously population and environment specific and are prone to misinterpretation (Visscher et al. 2008). Further community discussion on this concept is warranted.

Thermal responses in this experiment were also evaluated. These results indicate that there is high potential for thermal interactions with development time genotypes. In particular, these interactions are likely to affect other traits, including size, which is forensically informative. This observation would suggest that the genetic component to body size thermal responses may be more important than development time responses to temperature in the thermal ranges studied herein.

**Aim 2: Functional genetics of development of wild type strain**

**Collection and Treatments**

All lab work was done in the Texas A&M University Forensic Laboratory for Investigative Entomological Sciences facility. A wild type *Cochliomyia macellaria* population was collected from College Station, TX, in 2013 founded from >100 original individuals and resupplied every 2-3 generations during the course of the experiment. Blow fly eggs were collected estimating 200 eggs each. Once a time point of interest was reached (see Figure 12), samples were pulled from a jar, flash frozen for RNA isolation, and sampling ceased for that jar. Eggs were reared at four different temperature treatments (20 $^{O}$C, 25 $^{O}$C, 30 $^{O}$C and fluctuating) with 50% RH on a 14:10 L:D cycle.

**RNA Isolation and Sequencing**

The RNA-seq experiment was conducted as follows. Three biological replicates were used for each time point and each RNA sample was isolated from a pool of 5 individuals per life history stage using a standard Trizol (Invitrogen, Carlsbad, CA, USA) protocol. Samples were cleaned using a RNeasy Mini Kit (Qiagen) and quality checked using a



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

nanodrop (Thermo Scientific) and a BioAnalyzer (Agilent). A total of 84 RNA pools were sequenced on an Illumina HiSeq producing 100bp paired-end reads. Reads were then assembled into a transcriptome with a modification of the Oases algorithm (Schulz et al. 2012) (Table 10). Expression was assessed with eXpress (Roberts et al. 2011) and edgeR (Robinson et al. 2010) (Table 11). We have identified hundreds of genes that are differentially expressed between thermal treatments and 147 genes differentially expressed in a thermally fluctuating environment. Preliminary results have produced a list of hundreds of genes that are differentially expressed between stages with similar external morphologies. All comparisons have yielded significantly different expression of gene ontology groups (Tables 12-13).

## Future Directions, Forensic Implications, and Discussion

Our ability to collect preliminary data on *C. macellaria* development has enabled us to obtain internal funds from Texas A&M University to pursue expression of miRNA in wild type samples, as well as miRNA/mRNA expression in selected lines. These funds will enhance what we have already learned about our selected lines (providing functional genetic information about genes regulating development and development time variation). The RNA-seq project has produced a list of potential markers of developmental progress. We plan to confirm these in further studies and publication will provide a list that other groups can also test for usefulness in predicting blow fly age. Given the importance of protein metabolic processes identified here, it will also be interesting to pursue proteomic studies of blow fly development. Protein markers are expected to be more stable than mRNAs, which will enhance marker utility. Genomic and transcriptomic sequences will enable identification of any peptides identified in proteomic analyses.

## Summary

The funded project has already begun to shed further light on the genetics of blow fly development. It has expanded our knowledge of the role of genetics in development time variation, showing that there is ample wild genetic variation that could potentially impact forensic predictions. Our heritability estimates provide an empirical estimate of the impact of genetic variation on development time variation. The molecular biology projects in this proposal will enable us to pursue candidate genes that are markers of development time variation, developmental progress, and thermal exposure. Such candidates, studied by us or others that follow up on our results, can be developed into components of phenotype prediction and age prediction kits similar to the IrisPlex kit for human eye color prediction. It should be noted that at the moment our analyses are preliminary. Subsequent publications with the data presented in this report may differ from



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

future publications based on collection of further data, changes in parameter settings, differences in statistical tests performed, or choices in algorithms to apply to the data.

## Scholarly Products

Theses and Dissertations

1. Dissertation (Ph.D.) in Biochemistry and Biophysics at Texas A&M University. Shuhua Fu. Genomic and transcriptomic studies of non-model organisms. May 2015.
2. Thesis (M.S.) in Entomology at Texas A&M University. Ernesto Ramos III. A genetic study of the development of *Cochliomyia macellaria* (Fabricius) (Diptera: Calliphoridae): Ecological, evolutionary, and forensic importance of the secondary screwworm. August 2015.

Papers

1. SH Sze, AM Tarone. A memory-efficient algorithm to obtain splicing graphs and de novo expression estimates from de Bruijn graphs of RNA-seq data. BMC Genomics. 2015. 15 (Suppl 5), S6.
2. S Fu, AM Tarone, SH Sze. Heuristic pairwise alignment of de Bruijn graphs to facilitate simultaneous transcript discovery in related organisms from RNA-Seq data. BMC Genomics. 2015. In Press.

Book Chapters

1. AM Tarone. Chapter 14. Ecological Genetics. Carrion Evolution, Ecology, and Their Applications. Eds. ME Benbow, JK Tomberlin, AM Tarone. CRC Press. 2015.
2. AM Tarone, B Singh, CJ Picard. Chapter 24. Molecular Biology in Forensic Entomology. Forensic Entomology: International Dimensions and Frontiers. Eds. JK Tomberlin, ME Benbow. CRC Press. 2015.

Invited Presentations

1. AM Tarone presented on genomics in forensically important Diptera as part of a Workshop on Genomics (Moderated by Mary Curtis, US Fish and Wildlife National Forensic Laboratory) for the North American Forensic Entomology Association/Society for Wildlife Forensics Joint Meeting, Missoula, MT. June 24-26$^{th}$, 2015.
2. AM Tarone. CSI: Dipteran Genomics. USDA-ARS Knipling-Bushland U.S. Livestock Insects Research Laboratory, Kerrville, TX. December 9, 2014.
3. AM Tarone. CSI: Dipteran Genomics. University of North Texas Health Science Center, Center for Forensic Excellence. September 5, 2014.
4. AM Tarone. CSI: Dipteran Genomics. Purdue University, Department of Entomology. September 18, 2014.
5. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Selection for optimal phenotypes using genomics. Workshop on Insect Sensing, China Agricultural University, China, June 13$^{th}$, 2014.



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

6. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Selection for optimal phenotypes using genomics. Symposium of Sino-America on application of insect-microorganisms to treat organic waste materials. Zibo, China, June 12th, 2014.
7. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Selection for optimal phenotypes using genomics. Phoenix Black Soldier Fly Rearing Facility Workshop, Xi'an, China. June 11th, 2014.
8. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Genetic selection for optimal protein production and biodegradation phenotypes. Northwest Agricultural and Forestry University, Yangling, China. June 9th, 2014.
9. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Genetic selection for optimal protein production and biodegradation phenotypes. International Symposium on Organic Waste Bioconversion Mechanisms and Applications by Microbes and Insects, Wuhan, China. June 8th, 2014.
10. CJ Picard, AA Andere, E. Ramos, J Whale, J Parrott, AM Tarone. Genetic selection for optimal protein production and biodegradation phenotypes. Zheijiang University, Hangzhou, China June 2nd, 2014.
11. AM Tarone. Adventures in fly sex determination and life history trait evolution. University of Houston, Department of Biology and Biochemistry. April 2014.
12. C.J. Picard. Use of genomics in bridging basic and applied research areas of forensic entomology. Purdue University, Department of Entomology, February 2014.
13. C.J. Picard. Use of genomics in bridging basic and applied research areas of forensic entomology. Indiana University, Center for Computational Biology and Bioinformatics, November 2013.
14. AM Tarone, JK Tomberlin. Forensic Entomology at Texas State: What the Insects are Telling us. LBJ Student Center, Texas State University, San Marcos, TX, TX. Nov. 8, 2013.
15. C.J. Picard. Population Genomics in non-model organisms: RAD sequencing of a large sample of a forensically important blow fly. University of Dayton, Department of Biology, September 2013.

Conference Proceedings

1. JJ Parrott, E Ramos, C Spiegelman, ML Pimlsler, CJ Picard, AM Tarone. "Artificial selection on *Cochliomyia macellaria* (Fabricius; Diptera: Calliphoridae) development: Evolutionary Ecology and Forensic Implications. North American Forensic Entomology Association/Society for Wildlife Forensics Joint Meeting, Missoula, MT. June 24-26th, 2015.
2. AA Andere, E Ramos, J Parrott, J Whale, AM Tarone, CJ Picard. "Analysis of Genetic Variation in the Developmental Rate of the Blow Fly *Cochliomyia macellaria* (Diptera: Calliphoridae) based on their genomic sequences." North American Forensic Entomology Association/Society for Wildlife Forensics Joint Meeting, Missoula, MT. June 24-26th, 2015.



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

3. CG Owings, L Sercer, A Salamone, AM Tarone, CJ Picard. "Assessment of Genetic Variation Among Blow Fly Populations Exposed to Artificial Selection Pressures." North American Forensic Entomology Association/Society for Wildlife Forensics Joint Meeting, Missoula, MT. June 24-26th, 2015.
4. AA Andere, CG Owings, J Parrott, E. Ramos, AM Tarone, JW Whale, CJ Picard "Genetic variation in developmental time studied on genomic sequences of 3 geographically distinct populations of the blow fly *Cochliomyia macellaria* (Diptera: Calliphoridae)." 12th Annual Ecological Genomics Symposium. Kansas City, MO. October 31-November 2nd, 2014
5. AA Andere, CG Owings, J Parrott, E. Ramos, AM Tarone, J Whale, CJ Picard "Genetic variation in developmental time studied on genomic sequences of 3 geographically distinct populations of the blow fly *Cochliomyia macellaria* (Diptera: Calliphoridae). 12th Annual Ecological Genomics Symposium. Kansas City, MO.
6. CJ Picard "Predicting a phenotype from a genotype: Using carrion flies as a model organism to predict forensically relevant traits. Midwestern Association of Forensic Scientists, St. Paul, Minnesota October 8-10th, 2014.
7. CJ Picard, AA Andere, J Parrott, M Pimsler, E Ramos, AM Tarone, J Whale (2014) *How genomics is advancing the field of forensic entomology.* American Academy of Forensic Sciences (AAFS) Annual Meeting, Seattle, WA February 17-22nd.
8. E Ramos, CJ Picard, AM Tarone. Selecting for blow fly development: Forensically important *Cochliomyia macellaria*. Entomological Society of America Annual Meeting. Austin, TX, Nov. 12, 2013.
9. AM Tarone, LL Ellis, JS Johnston, CJ Picard. Consequences of genome size variation in forensic entomology. Entomological Society of America Annual Meeting. Austin, TX, Nov. 12, 2013.
10. AM Tarone. The genomics of blow fly development: Advancing research in forensic science and sex determination. Texas Genetics Society. College Station, TX. May 2013.
11. SH Sze, AM Tarone. A memory-efficient algorithm to obtain splicing graphs and de novo expression estimates from de Bruijn graphs of RNA-seq data. 3rd Workshop on Computational Advances for Next Generation Sequencing (CANGS'2013) at the 3rd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS'2013). June 12-14, New Orleans, LA.

## Expected Scholarly Products

We expect to submit manuscripts for publication on the following topics: The phenotypic response to selection, the thermal plasticity of selected flies, selection genomics, candidate gene responses to selection, transcriptome expression over time, miRNA expression over time, miRNA/mRNA in selected lines, and transcriptomic assembly algorithms.



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

## Literature Cited


**Blanckenhorn, W. U. 1997.** Altitudinal life history variation in the dung flies *Scathophaga stercoraria* and *Sepsis cynipsea*. Oecologia 109: 342-352.

**Blanckenhorn, W. U. 1998.** Adaptive phenotypic plasticity in growth, development, and body size in the yellow dung fly. Evolution 52: 1394-1407.

**Byrd, J., and J. Castner (eds.). 2010.** Forensic entomology: The utility of arthropods in legal investigations. CRC Press, Boca Raton, FL.

**Conner, J. K., and D. L. Hartl. 2004.** A primer of ecological genetics, Sinauer Associates, Sunderland, Mass.

**Falconer, D. S. and Mackay T.F.C. 1996.** Introduction to quantitative genetics, 4th ed. Longman, New York, NY.

**Gallagher, M. B., S. Sandhu, and R. Kimsey. 2010.** Variation in developmental time for geographically Distinct populations of the common green bottle fly, *Lucilia sericata* (Meigen). Journal of Forensic Sciences 55: 438-442.

**Mousseau, T. A., and D. A. Roff. 1987.** Natural selection and the heritability of gitness components. Heredity 59: 181-197.

**Mousseau, T. A., and H. Dingle. 1991.** Maternal effects in insect life histories. Annual Review of Entomology 36: 511-534.

**Owings, C. G., C. Spiegelman, A. M. Tarone, and J. K. Tomberlin. 2014.** Developmental variation among *Cochliomyia macellaria* Fabricius (Diptera: Calliphoridae) populations from three ecoregions of Texas, USA. International Journal of Legal Medicine 128: 709-717.

**Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. 2011.** Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology 12.

**Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010.** edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

**Roff, D. A., and T. A. Mousseau. 1987.** Quantitative genetics and fitness - Lessons from *Drosophila*. Heredity 58: 103-118.

**Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney. 2012.** Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086-1092.

**Tarone, A. M., and D. R. Foran. 2008.** Generalized additive models and *Lucilia sericata* growth: Assessing confidence intervals and error rates in forensic entomology. Journal of Forensic Sciences 53: 942-948.

**Tarone, A. M., C. J. Picard, C. Spiegelman, and D. R. Foran. 2011.** Population and temperature effects on *Lucilia sericata* (Diptera: Calliphoridae) body size and minimum development time. Journal of Medical Entomology 48: 1062-1068.

**Tomberlin, J. K., M. E. Benbow, A. M. Tarone, and R. M. Mohr. 2011a.** Basic research in evolution and ecology enhances forensics. Trends in Ecology and Evolution 26: 53-55.

**Tomberlin, J. K., R. Mohr, M. E. Benbow, A. M. Tarone, and S. VanLaerhoven. 2011b.** A roadmap for bridging basic and applied research in forensic entomology. Annual Review of Entomology 56: 401-421.

**VanLaerhoven, S. L. 2008.** Blind validation of postmortem interval estimates using developmental rates of blow flies. Forensic Science International 180: 76-80.

**Visscher, P. M., W. G. Hill, and N. R. Wray. 2008.** Heritability in the genomics era: Concepts and misconceptions. Nature Reviews Genetics 9: 255-266.




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

## Appendix

**Table 1. Mean development times for first and 23rd generation of selection experiments.**

**Replicate 1**

| Population | G01 Mean | G23 Slow Mean | G23 Fast Mean | Δ G23 |
|---|---|---|---|---|
| College Station | 259 | 419 | 245 | 174 |
| Longview | 284 | 367 | 247 | 120 |
| Snook | 273 | 307 | 237 | 70 |

**Replicate 2**

| Population | G01 Mean | G23 Control Mean | G23 Slow Mean | G23 Fast Mean | Δ G23 |
|---|---|---|---|---|---|
| College Station | 261 | 254 | 355 | 223 | 132 |
| Longview | 269 | 295 | 357 | 218 | 139 |
| San Marcos | 278 | 279 | 345 | 217 | 128 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 2. Example ANOVA results for both replicates of selection. This model represents a likely model for the selection response based on AICc comparisons of possible models.**

| Replicate 1 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Phenotype | Model | Source | Df | Sum of Squares | Mean Square | F ratio | Prob.>F |
| Development | ANOVA | Model | 47 | 218104.70 | 4640.53 | 16.94 | <.0001 |
| | | Error | 90 | 24652.45 | 273.92 | | |
| | | Total | 137 | 242757.14 | | | |
| | Effect tests | Generation x Selection | 22 | 45361.64 | | 7.53 | <.0001 |
| | | Generation | 22 | 21825.93 | | 3.62 | <.0001 |
| | | Selection | 1 | 142283.86 | | 519.44 | <.0001 |
| | | Population | 2 | 8633.28 | | 15.76 | <.0001 |
| **Replicate 2** | | | | | | | |
| Phenotype | Model | Source | Df | Sum of Squares | Mean Square | F ratio | Prob.>F |
| Development | ANOVA | Model | 7 | 99569.87 | 14224.30 | 334.71 | <.0001 |
| | | Error | 130 | 5524.61 | 42.50 | | |
| | | Total | 137 | 105094.47 | | | |
| | Effect Tests | Generation x Selection | 22 | 27053.36 | | 318.30 | <.0001 |
| | | Generation | 22 | 138.81 | | 3.27 | 0.073 |
| | | Selection | 1 | 64375.69 | | 757.41 | <.0001 |
| | | Population | 2 | 4426.93 | | 52.09 | <.0001 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 3: Libraries and genome assemblies generated.**

| Geographic Location | Selection | Genome Label |
|---|---|---|
| Longview, TX | None, original baseline population | L-B |
| Longview, TX | Fast Selection (generation 26) | L-F |
| Longview, TX | Slow Selection (generation 26) | L-S |
| Snook, TX | None, original baseline population | S-B |
| Snook, TX | Fast Selection (generation 26) | S-F |
| Snook, TX | Slow Selection (generation 26) | S-S |
| College Station Airport, TX | None, original baseline population | A-B |
| College Station Airport, TX | Fast Selection (generation 26) | A-F |
| College Station Airport, TX | Slow Selection (generation 26) | A-S |
| Combined Baseline Populations | None | Cmbd-B |
| Combined FAST Populations | Fast Selection | Cmbd-F |
| Combined SLOW Populations | Slow Selection | Cmbd-S |

**Table 4: Summary statistics for 9 draft genome assemblies.**

| Genome | Kmer (bp) | # reads (millions) | % reads mapped | # contigs | N50 (bp) | Average Length (bp) | Max Length (bp) | Genome Size (Mbp) |
|---|---|---|---|---|---|---|---|---|
| L-B | 32 | 78.3 | 58.0 | 143,437 | 1,238 | 1,186 | 63,186 | 170 |
| L-F | 32 | 77.1 | 86.0 | 225,322 | 1,614 | 1,398 | 1,254,171 | 314 |
| L-S | 32 | 79.2 | 85.2 | 227,338 | 2,050 | 1,651 | 502,359 | 375 |
| C-B | 32 | 85.3 | 70.0 | 394,242 | 850 | 645 | 110,440 | 254 |
| C-F | 32 | 94.3 | 81.9 | 260,371 | 1,648 | 1,391 | 1,254,024 | 362 |
| C-S | 32 | 66.1 | 79.8 | 212,794 | 1,377 | 1,300 | 1,567,063 | 276 |
| Cmbd-B | 32 | 233.3 | 63.9 | 166,283 | 1,476 | 1,332 | 106,709 | 221 |
| Cmbd-F | 25 | 269.7 | 83.4 | 444,662 | 1,402 | 1,234 | 1,571,258 | 548 |
| Cmbd-S | 32 | 239.2 | 86.1 | 375,920 | 1,759 | 1,446 | 1,096,330 | 543 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 5: List of *Drosophila* candidate genes and results seen in coding sequences in the fast and slow draft genomes.**

| Gene name | Drosophila accession number | Results – coding sequences only |
|---|---|---|
| Giant (gt) | CG7952 | Gene predicted only in slow genome |
| Forkhead | CG10002 | Gene predicted only in fast genome |
| Pten | CG5671 | Gene predicted only in fast genome |
| Torso (tor) | CG1389 | Gene predicted only in slow genome |
| Egfr | CG10079 | Identical gene structures in fast and slow genomes |
| Happyhour | CG7097 | 1 SNP, non-synonymous mutation (L/I amino acid change) |
| bitesize | CG44012 | Multiple non-synonymous SNPs and 3bp indels, |
| Diminutive (Dm) | CG10798 | Gene predicted only in slow genome |
| InR | CG18402 | Multiple synonymous SNPs, 1 indel |
| Minus (mi) | CG5360 | Gene predicted only in slow genome |
| Neurofibromin 1 (Nf1) | CG8318 | Multiple synonymous SNPs |
| Short neuropeptide F receptor (sNPF-R) | CG7395 | Multiple synonymous SNPs |
| Tctp | CG4800 | Identical coding sequences (1 large intron) |
| Target of rapamycin (Tor) | CG5092 | Multiple SNPs, a large portion of amino acids are missing (~25 aa) as well as additional structural changes |

**Table 6: List of loci initially explored as candidates for a molecular marker of fast and slow development.**

| Primer Names | Polymorphisms predicted | Expected Product Sizes | |
|---|---|---|---|
| | | Slow Genome | Fast Genome |
| Btsz_1 | 3 SNPs | 965 bp | 965 bp |
| InR_1 | 5 SNPs; 1 Indel (3bp) | 573 bp | 576 bp |
| InR_2 | 7 SNPs; 1 Indel (11bp) | 688 bp | 677 bp |
| Tctp_1 | 8 SNPs; 2 Indels (112bp & 3bp) | 806 bp | 691 bp |
| Tor_1 | 3 SNPs; 1 Indel (104bp) | 460 bp | 564 bp |
| Tor_2 | 6 SNPs | 599 bp | 599 bp |

**Table 7: Mapping statistics from each individual library to the SLOW combined (Cmbd-S) genome.**

| Selection | Parameters | College Station | Longview | Snook |
|---|---|---|---|---|
| None (Baseline) | # reads (millions) | 85.3 | 84.0 | 82.3 |
| | % mapped | 74.4 | 74.5 | 74.3 |
| Fast | # reads (millions) | 99.4 | 80.6 | 104.5 |
| | % mapped | 77.8 | 82.3 | 76.1 |
| Slow | # reads (millions) | 68.1 | 83.5 | 99.6 |
| | % mapped | 83.2 | 82.6 | 84.3 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 8: summary statistics of the variant distribution across all individual populations when mapped back to either the baseline, fast or slow genomes (combined). Variants detected include SNPs (single nucleotide polymorphisms), MNVs (multiple nucleotide polymorphisms), insertions and deletions. Filtering parameters included a read balance (of forward/reverse reads) of ≥0.4 and a minimum coverage of 20 reads.**

| Selection | Parameters | College Station | Longview | Snook |
|---|---|---|---|---|
| None (Baseline) | # Variants (raw) | 3,735,069 | 3,806,614 | 3,535,145 |
| | # Variants (filtered) | 67,220 | 63,929 | 46,984 |
| | SNPs | 53,004 | 50,301 | 37,325 |
| | MNVs | 1,759 | 1,732 | 1,286 |
| | Insertions | 6,415 | 6,149 | 4,255 |
| | Deletions | 6,042 | 5,747 | 4,118 |
| Fast | # Variants (raw) | 4,364,835 | 2,395,867 | 5,033,861 |
| | # Variants (filtered) | 65,272 | 45,852 | 155,033 |
| | SNPs | 51,314 | 42,873 | 118,558 |
| | MNVs | 1,928 | 1,130 | 4,650 |
| | Insertions | 6,142 | 843 | 16,267 |
| | Deletions | 5,888 | 1,006 | 15,558 |
| Slow | # Variants (raw) | 1,428,258 | 2,395,042 | 2,489,738 |
| | # Variants (filtered) | 46,825 | 51,125 | 39,992 |
| | SNPs | 44,449 | 47,568 | 36,081 |
| | MNVs | 1,044 | 1,509 | 1,101 |
| | Insertions | 594 | 954 | 1,337 |
| | Deletions | 738 | 1,094 | 1,473 |

**Table 9: Zygosity distribution of single nucleotide polymorphisms in each draft genome. The Snook population had many more SNPs than any of the other two populations, however, after filtering; many of these were eliminated (low coverage SNPs).**

| Genome | Heterozygous SNPs | Homozygous SNPs |
|---|---|---|
| L-B | 48,603 | 1,698 |
| L-F | 39,349 | 3,524 |
| L-S | 45,896 | 1,672 |
| S-B | 36,048 | 1,277 |
| C-B | 51,177 | 1,827 |
| C-F | 45,332 | 5,992 |
| C-S | 36,549 | 7,900 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 10. Assembly statistics for our computational pipeline with model organisms and for *C. macellaria*. Reported are the number of estimated loci, number of estimated transcripts (parenthetical value is ration of transcripts to loci, as a genetic locus can have more than one transcript due to alternative splicing), n50 (median transcript size), and BLASTX results (numbers of loci with BLAST hits to *D. melanogaster* genes).**

| #Loci | #Transcripts | n50 | #BLASTX |
|---|---|---|---|
| 41159 | 60812 (1.48) | 2223 | 22370 |

**Table 11. Gene expression differentiation in *C. macellaria* by temperature and by developmental stage as determined by analyzing predicted transcripts with eXpress and edgeR. Temperatures were designated by T followed by the degrees Celsius for growth at constant temperatures or Flux for fluctuations between 20ºC and 30ºC that averaged to 25ºC. Developmental progress was compared among feeding third intars (F3I), early postfeeding third instars (EPF), late postfeeding third instars (LPF), early pupae (EP), mid-pupae ~1/3 through development (MP1), mid-pupae ~2/3 through development (MP2), and late pupae (LP).**

| Temperature | Differentially Expressed Genes |
|---|---|
| T20vsT25 | 361 |
| T20vsT30 | 369 |
| T20vsFlux | 415 |
| T25vsT30 | 511 |
| T25vsFlux | 147 |
| T30vsFlux | 415 |

| Development | Differentially Expressed Genes |
|---|---|
| F3IvsEPF | 288 |
| EPFvsLPF | 367 |
| LPFvsEP | 1563 |
| EPvsMP1 | 1315 |
| MP1vsMP2 | 984 |
| MP2vsLP | 1091 |

**Table 12. Gene ontology categories significantly enriched in the sets of genes differentially expressed by temperature between T25 and Flux samples. These are putative markers of fluctuating temperatures.**

| GO Category | Number of Genes in Category |
|---|---|
| Oxidation-reduction process | 107 |
| Single-organism metabolic process | 134 |
| Secondary metabolic process | 26 |
| Secondary metabolite biosynthetic process | 16 |
| Response to insecticide | 14 |
| Melanin biosynthetic process | 11 |
| Hormone metabolic process | 17 |
| Response to toxic substance | 14 |
| Phenol-containing compound biosynthetic process | 11 |
| Steroid metabolic process | 14 |

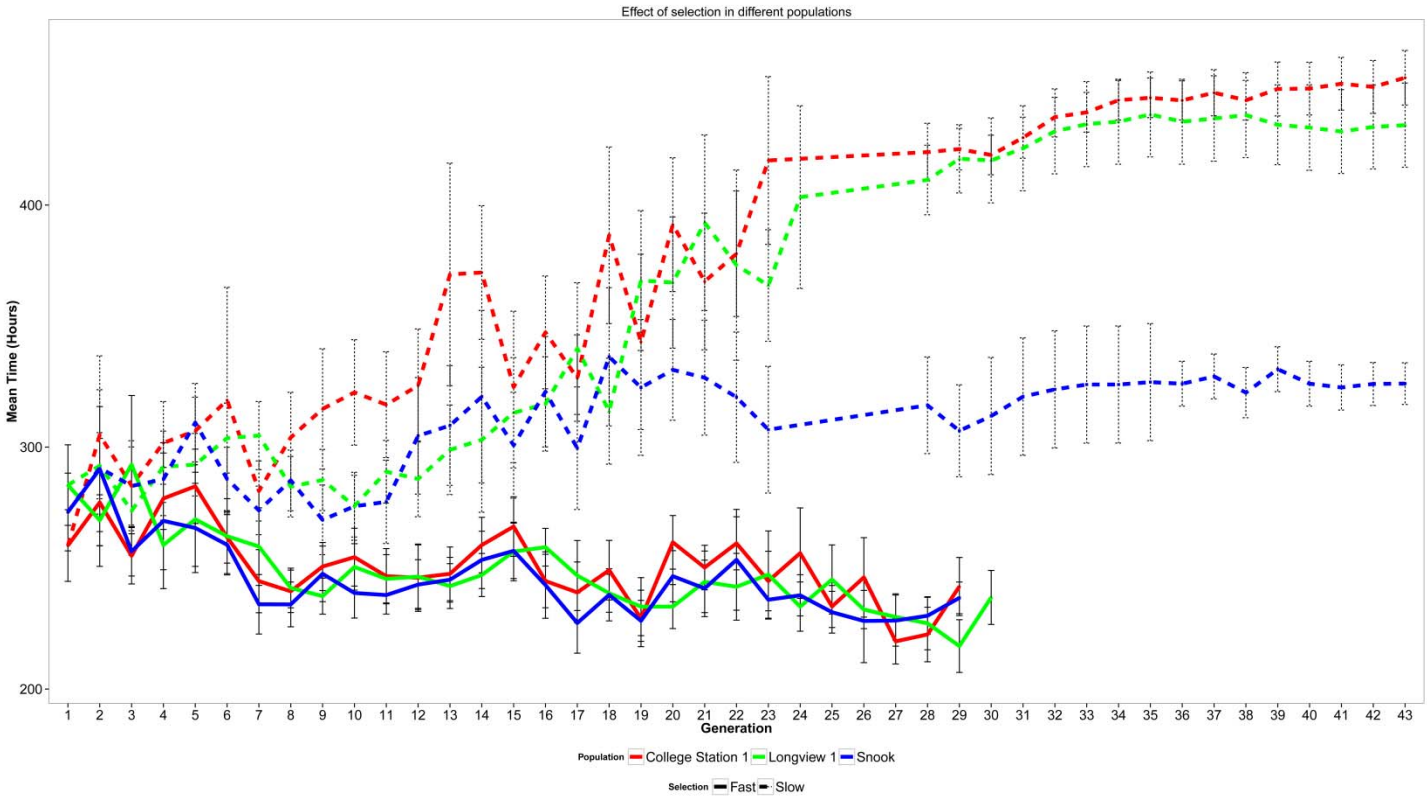
This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Table 13. Gene ontology categories significantly enriched in the sets of genes differentially expressed between early and late postfeeding third instars. These are candidate markers of developmental progress.**

| GO Category | Number of Genes in Category |
|---|---|
| Proteolysis | 252 |
| Protein metabolic process | 419 |
| Protein phosphorylation | 144 |
| Phosphorylation | 145 |
| Metabolic process | 641 |
| Primary metabolic process | 534 |
| Macromolecule metabolic process | 477 |
| Transmembrane transport | 102 |
| Phosphorus metabolic process | 159 |
| Phosphate-containing compound metabolic process | 157 |

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 1. Development times during the first replicate of the selection experiment. Lines connect means within a selection group over time. Bars represent standard deviations per generation. Colors indicate source populations. Fast populations went extinct between generations 29 and 30. Slow populations are extant after 43 generations.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 2. Development times during the second replicate of selection. Figure details are as in Figure 1. Note the addition of controls. No population has gone extinct after 23-26 generations of selection.**

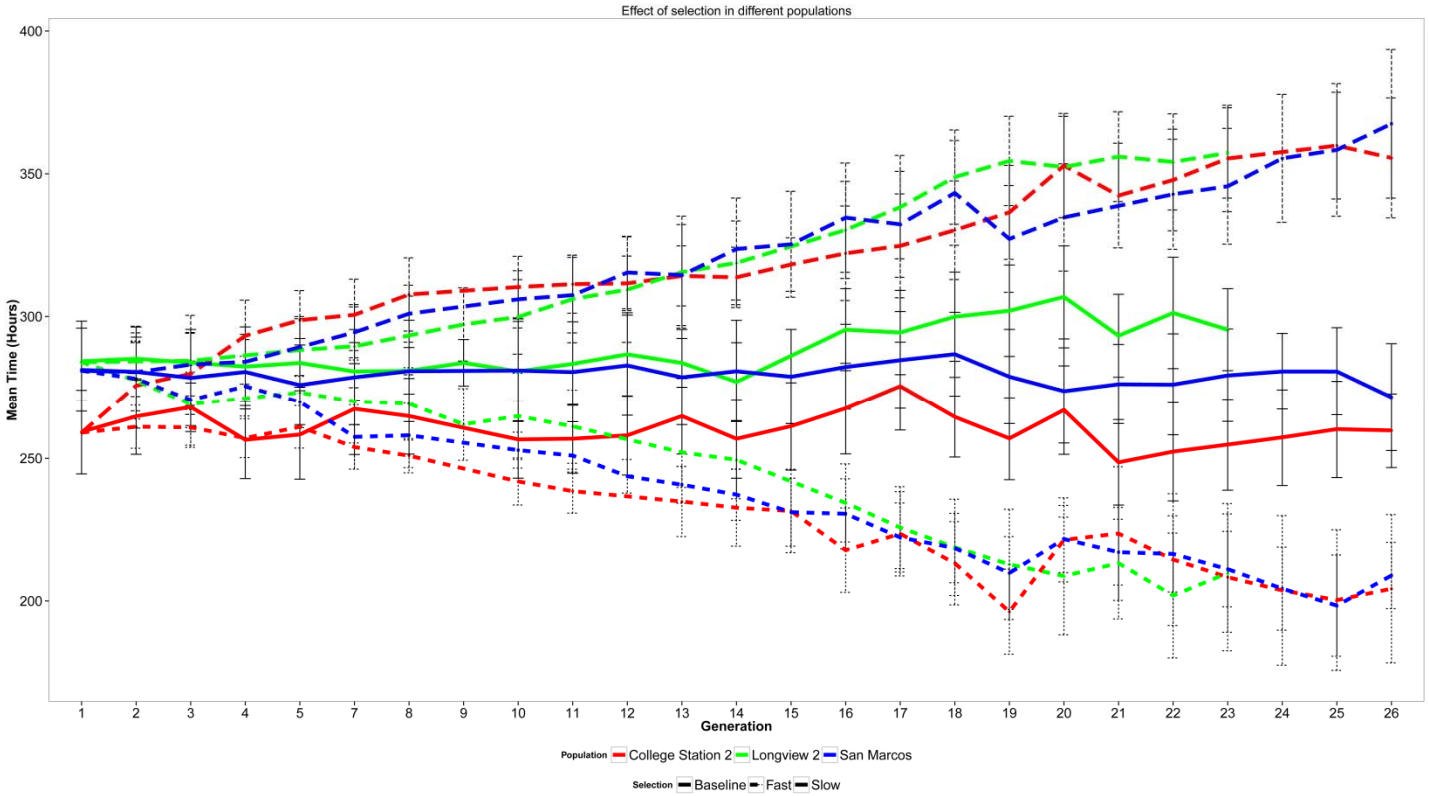
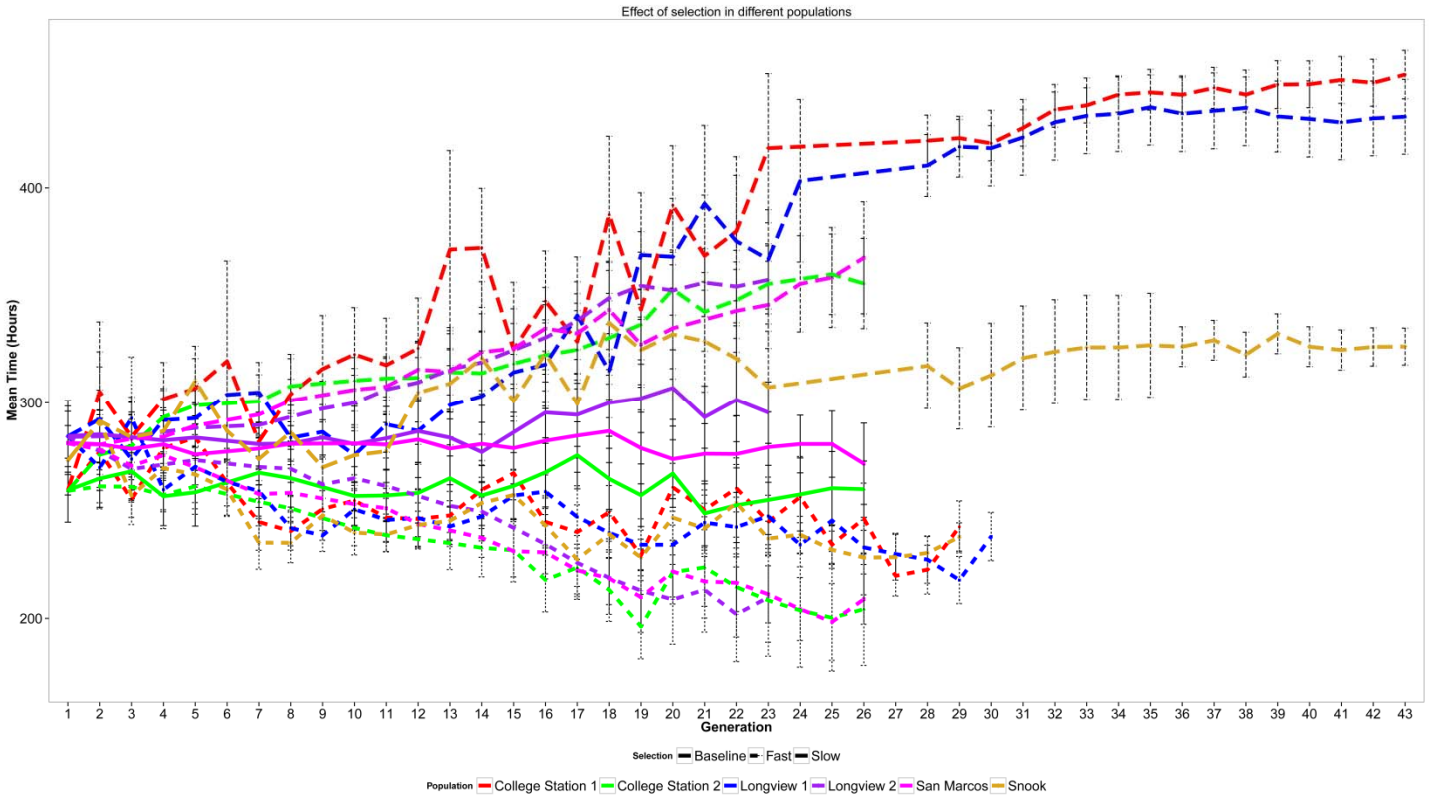
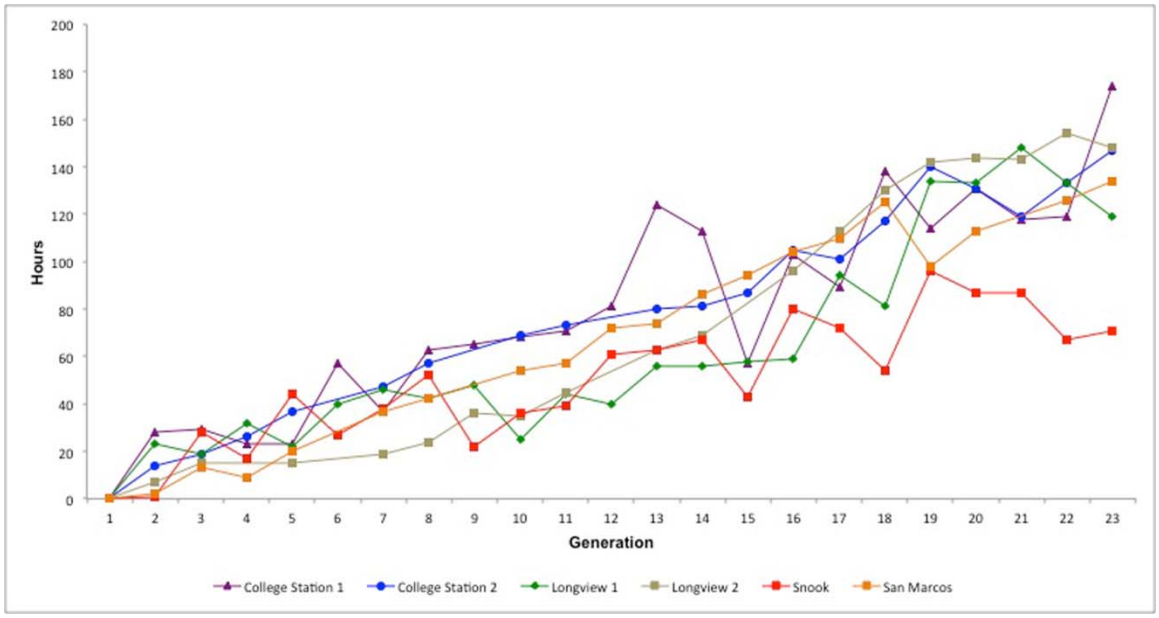



**Figure 3. Combined selection responses of all strains studied. Figure details are as in Figure 1.**



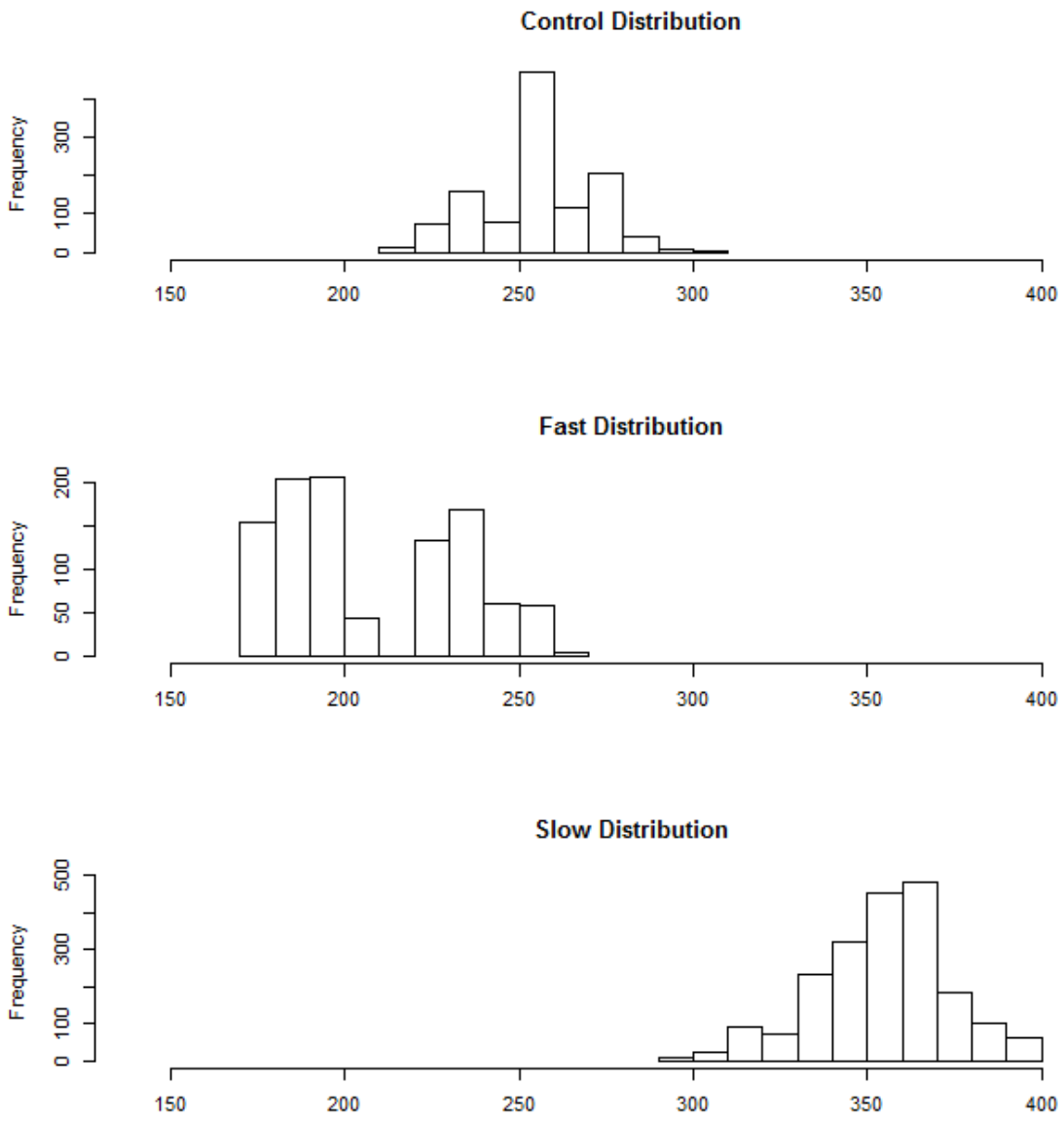

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 4. Selection differential (mean development time of slow selection group minus the mean development time for the fast selection group) for all 6 populations.**



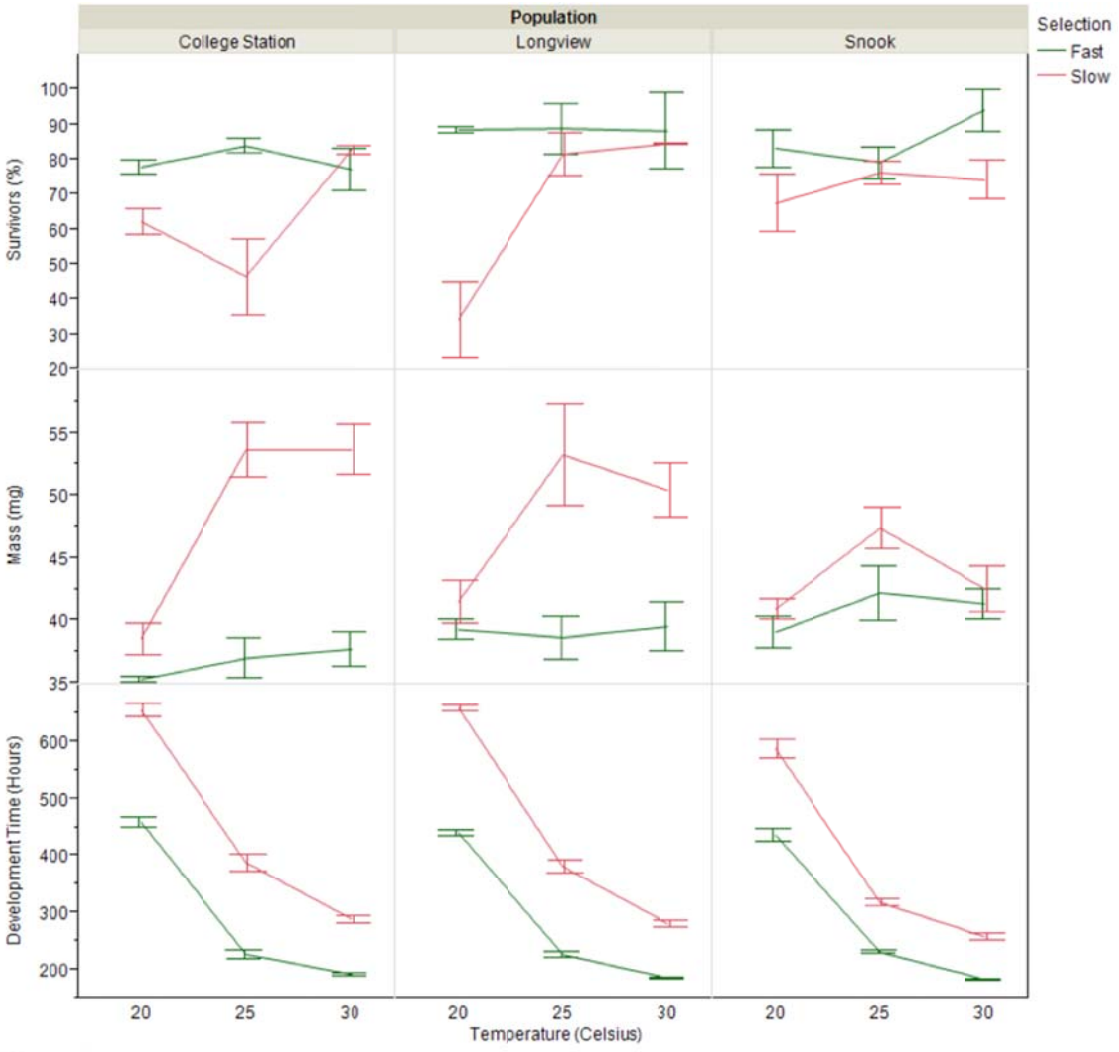

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 5. Distributions of development time in hours for each selection group in the College Station strains of the second replicate of selection. Results are from the 23rd generation of the experiment.**



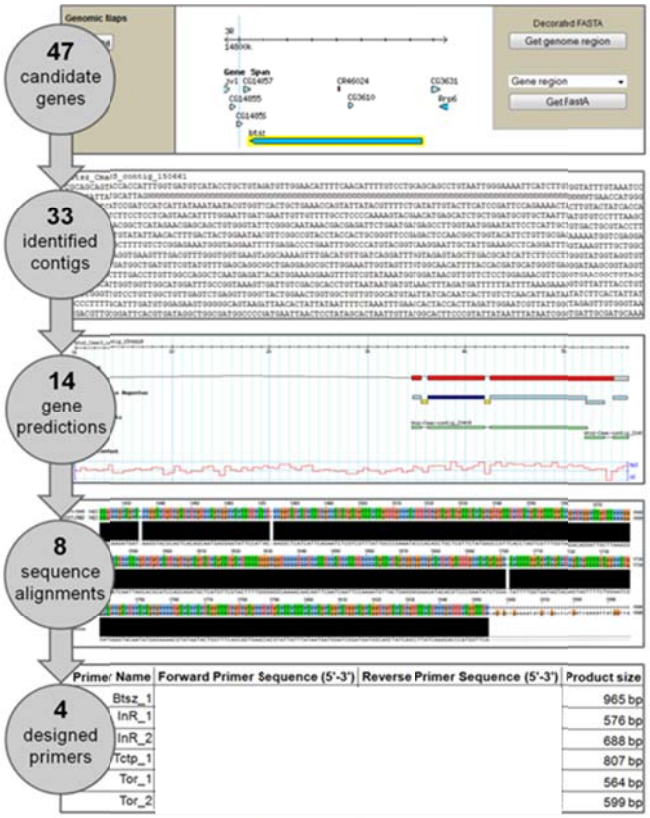

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 6. Development times, pupal masses, and immature survivorship across all three populations and for both selection groups. Note the presence of thermal, selection group, and population specific responses. Bars represent standard errors of phenotype means.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 7: Pipeline for candidate gene approach.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 8: Example of gene prediction in slow and fast combined genomes using *C. macellaria* transcripts as a constraint on gene prediction.**

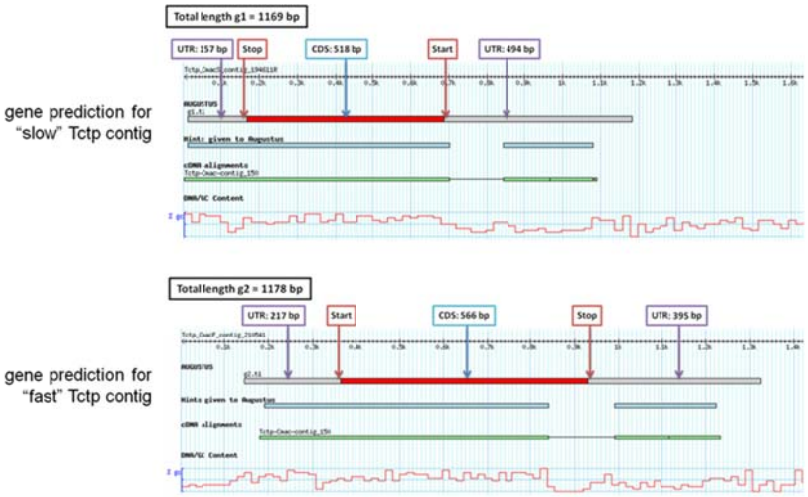
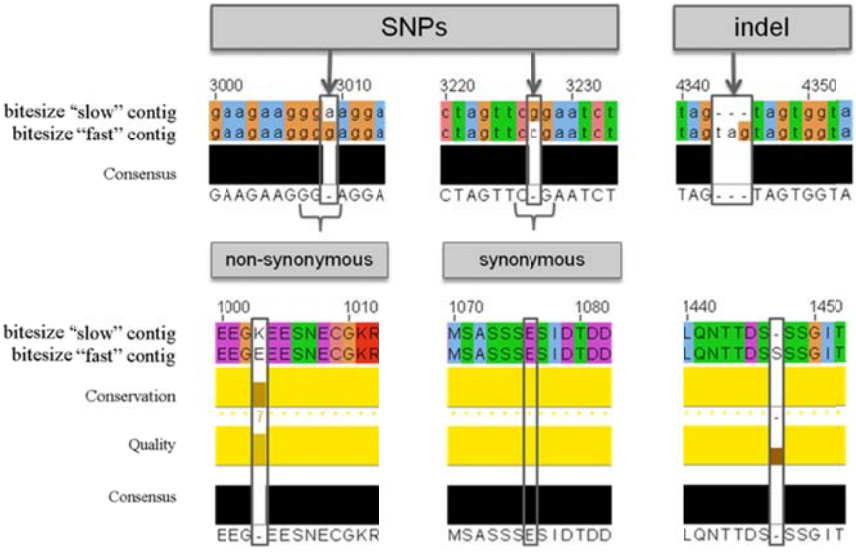



**Figure 9: Example alignments of slow and fast contigs demonstrated synonymous, non-synonymous and indels detected in our candidate gene approach.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 10: Gel image of PCR amplification of the InR_2 locus using Longview slow (lane 1) and fast (lane 2) specimens that were not a part of the original sequencing cohort.**

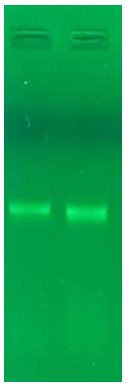
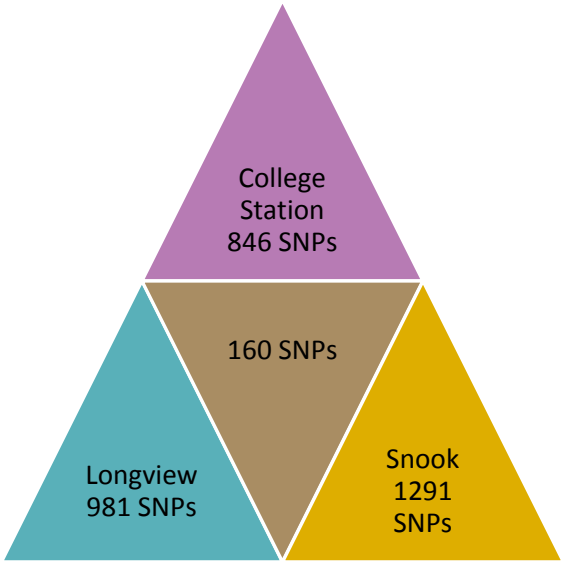
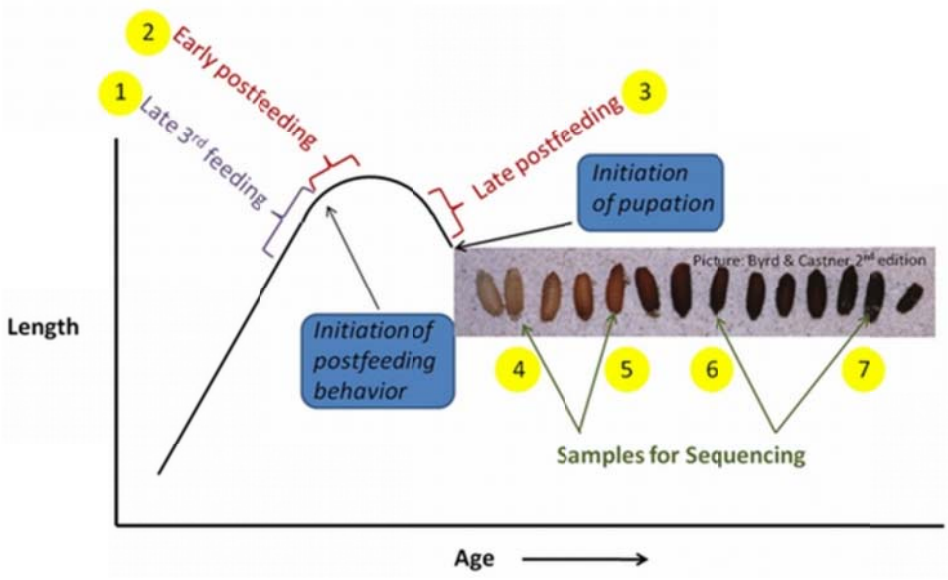


**Figure 11: Comparative analyses of differentially fixed SNPs (between fast and slow groups within a population) present in all three of the individual libraries (baseline, fast and slow) for each region, with a total of 160 SNPs present across all nine draft genomes.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

**Figure 12. Schematic for sampling design in the RNA-seq experiment.**




This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.