Promoter Identification in *Daphnia* Populations Revealed by Transcription Start Site Profiling

by

Shannon Snyder

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2020 by the
Graduate Supervisory Committee:

Michael Lynch, Chair
Robin Harris
Randolph Raborn
Jeremy Wideman

ARIZONA STATE UNIVERSITY

August 2020

ABSTRACT

Regulation of transcription initiation is a critical factor in the emergence of diverse biological phenotypes, including the development of multiple cell types from a single genotype, the ability of organisms to respond to environmental cues, and the rise of heritable diseases. Transcription initiation is regulated in large part by promoter regions of DNA. The identification and characterization of *cis*-regulatory regions, and understanding how these sequences differ across species, is a question of interest in evolution. To address this topic, I used the model organism *Daphnia pulex*, a well-characterized microcrustacean with an annotated genome sequence and selected a distribution of well-defined populations geographically located throughout the Midwestern US, Oregon, and Canada. Using isolated total RNA from adult, female *Daphnia* originating from the selected populations as well as a related taxon, *Daphnia pulicaria* (200,000 years diverged from D. *pulex*), I identified an average of over 14,000 (n=14,471) promoter regions using a novel transcription start site (TSS) profiling method, STRIPE-seq. Through the identification of sequence architecture, promoter class, conservation, and transcription start region (TSR) width, of *cis*-regulatory regions across the aforementioned *Daphnia* populations, I constructed a system for the study of promoter evolution, enabling a robust interpretation of promoter evolution in the context of the population-genetic environment. The methodology presented, coupled with the generated dataset, provides a foundation for the study of the evolution of promoters across both species and populations.

ACKNOWLEDGMENTS

This work was made possible through a grant (#5R35GM122566) from the NIH National Institute for General Medical Sciences awarded to ML. Development of STRIPE-seq was partially supported by seed funding to RTR from the Indiana Clinical and Translational Sciences Institute funded in part by Grant Number UL1TR001108. from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. The authors acknowledge the Indiana University Pervasive Technology Institute and ASU Research Computing for providing HPC resources that have contributed to the research results.

Additionally, I would like to thank Michael Lynch for his gracious support; Emily Williams for her abundant teaching, and dedication to the *Daphnia*; R. Taylor Raborn for his persistent guidance, unwavering encouragement, and considerate mentorship throughout the completion of this project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

Term	Abbreviation
1. Base Pairs	bp
2. B Recognition Element	BRE
3. BRE Downstream	BREd
4. BRE Upstream	BREu
5. Complementary DNA	cDNA
6. Cap Analysis of Gene Expression	CAGE
7. Chromatin Immunoprecipitation	ChIP
8. Downstream Core Promoter Element	DCE
9. Empirical Analysis of Digital Gene Expression Data in R	edgeR
10. General Transcription Factors	GTF
11. Initiator Motif	Inr
12. Modified Shape Index	MSI
13. Motif 10 Element	MTE
14. Reverse Transcription Oligonucleotide	RTO
15. Ribonucleic acid	RNA
16. RNA Annotation and Mapping of Promoters for the Analysis of Gene Expression …	RAMPAGE
17. Serial Analysis of Gene Expression	SAGE
18. Shape Index	SI
19. Survey of Transcription Initiation and Promoters Elements	STRIPE-seq
20. TATA Box-binding Protein	TBP

Term Abbreviation

21. Template Switching Oligonucleotide ..... TSO

22. Template Switching Reverse Transcription ..... TSRT

23. Transcription preinitiation complex ..... PIC

24. Transcription Start Region ..... TSR

25. Transcription Start Site ..... TSS

26. Untranslated region ..... UTR

CHAPTER 1

INTRODUCTION

Transcription Initiation

Gene expression is a fundamental process that underlies all of biology, namely through the regulation of developmental pathways, the maintenance of homeostasis, the divergence of diverse phenotypes, and the onset of disease (Vacik, et al, 2017). Gene expression is dependent on the synchronized assembly of mRNA transcripts (Lenhard et al., 2012). In eukaryotes, the multiprotein RNA polymerase II complex associates with the core promoter to synthesize mRNA from the DNA template (Haberle and Stark, 2018). The core promoter is traditionally described as the minimal stretch of DNA required to facilitate transcription initiation through the association with RNA polymerase II (Sandelin et al., 2007). The transcription start site (TSS) is defined as the first nucleotide of the synthesized mRNA, which defines the [+1] position of the core promoter region.

In order to bind to the core promoter, RNA polymerase II requires the localization of general transcription factors (GTFs) at the promoter sequence (Lenhard et al., 2012). Among these GTFs is TFIIA, which includes the TATA box-binding protein and TAFs or TBP-associated factors (Juven-Gershon et al., 2008). Other associated basal transcription factors included in the GTF family are TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (Kadonaga, 2012). The core promoter is defined as the minimal stretch of DNA required to facilitate transcription, and is comprised of a number of sequence motifs, known as core promoter elements. One of these promoter elements is BRE, which acts as a transcription factor recognition element, as well as the TATA box which is only present

in 10-20% of metazoan genomes. The TATA box is located roughly 25 to 30 base pairs upstream of the TS [-25 to -30]. Other core promoter elements include the Initiator (Inr) motif found at the start site (positions -2 to +4), and the Downstream Core Promoter Element or DPE (positions +28 to +32) (Juven-Gershon et al., 2008; Ponjavic et al., 2006; Sandelin et al., 2007). During the process of transcription initiation, the GTFs and RNA polymerase II associate at the core promoter to form the transcription preinitiation complex or PIC. From here, transcription is initiated, and the resulting transcript begins the elongation phase on the template DNA (Kadonaga, 2012).

Transcription initiation and its far reaching implications is also evident in the use of alternative promoters, in which various alternative transcripts can be generated attributing to transcriptomic diversity. This occurrence further demonstrates the importance of promoters and their control of gene regulation especially due to its role in numerous developmental defects and disease, cancer, schizophrenia, and Parkinson's disease (Davuluri et al., 2008; Vacik & Raska, 2017; Demircioğlu et al., 2019). Additionally, there is evidence of alternative promoter usage taking place across a range of eukaryotic taxa including humans, mice, and plants (Gregory, 2018; Reyes & Huber, 2018; Zhang et al., 2007), which commonly contributes to transcripts with differing 5`UTR regions, leading to distinct protein isoforms, and, in certain cases, functional outcomes.

Identification of Promoters

In vitro analysis. Prior to the introduction of high-throughput, genome-wide promoter identification assays, common methods of promoter characterization

incorporated gene-by-gene analysis and various molecular biology cloning techniques including nuclease protection and primer extension (Sandelin et al., 2007).

Nuclease Protection Assay. One such method is the nuclease protection assay, wherein a DNA probe containing the predicted TSS is hybridized to an mRNA fragment. The hybrid molecule is digested with a nuclease (commonly S1)(Sandelin et al., 2007) which degrades single-stranded genomic fragments; any genomic materials that do not contain both the mRNA and the DNA probe hybrid are digested and removed from the assay. Following subsequent gel electrophoresis, the length of the resulting molecule can allow for the detection of the TSS according to its size (Sandelin et al., 2007).

Primer Extension. Another gene specific method, primer extension, involves the incorporation of a labeled primer, which is complementary to the mRNA to the 5`-end of the molecule of interest (Sandelin et al., 2007). From there, reverse transcription synthesizes cDNA from the point the primer anneals to the end of the mRNA template. The distance of the extension of the primer product corresponds with the length to the corresponding TSS, as is the case with the nuclease protection assay.

Reporter gene assays. Other methods of gene-specific TSS identification commonly involve the cloning of genomic sequence coupled with the use of a reporter gene to determine functionality. Another approach in identifying transcriptionally relevant stretches of DNA involve the systematic deletion of nucleotides to determine crucial regulatory elements, while others identify key promoter elements, such as TATA or transcription factor binding sites using computational methods.

Rapid amplification of 5` complementary DNA ends (5` RACE). 5` RACE is a reverse transcription based technique which is utilized to extend and amplify partial

cDNA clones. 5` RACE amplifies the 5` ends of the corresponding mRNA molecule through the addition of a homopolymeric tail, which provides a binding platform upstream of the target mRNA and aids in the identification of the 5` region, and thus the transcription start site ("Rapid amplification of 5′ complementary DNA ends (5′ RACE)," 2005).

The major drawback of the aforementioned gene-specific methods is they are of low to moderate throughput. As such, none of the approaches are well-suited for genome-wide analysis. They are also highly laborious and time-consuming, which is especially apparent in cloning assays which also require promoter function verification in addition to the construction and isolation of the genomic sequence. Many of these identification methods, including nuclease protection assay, require an interpretation of fragment size following gel-electrophoresis which introduces an additional source of uncertainty and reduced resolution.

Hybridization and Functional Genomics Methods for Promoter Identification

Branching from the localized approaches at promoter identification and annotation are the functional genomic methods. Functional genomics describes the field of biology focused on genome-wide approaches to addressing the link between the genotype and phenotype, with a major goal focused on the annotation of genomic regions. Several consortiums, including The Encyclopedia of DNA Elements (ENCODE) and Functional Annotation of the Mammalian Genome (FANTOM) have been assembled with a main objective of functional annotation. The vast availability of genome and transcriptome data, much of which was generated by functional genomics methods

(including RNA-seq (Lister et al., 2008), ChIP-seq (Nelson et al., 2006), and DNAse I hypersensitivity (Pipkin & Lichtenheld, 2006)(Le Roch, 2013) has paved the way for investigations into the annotation of functional regions, including the identification and investigations of promoters and their corresponding characteristics. Large-scale approaches which utilize the availability of well-annotated reference assemblies have underpinned characterization of promoter location on a genome-wide scale, rather than on a gene-by-gene basis. Some of these approaches are dependent on the identification of the 5`-ends of mRNA transcripts, and the utilization of reverse transcription to construct a complete cDNA library. A common distinction concerns the use of full-length cDNA tags (paired-end), or isolation of the 5` end of the DNA reads (single-end); paired-end sequencing involves obtaining sequences from two distinct parts of a DNA molecule, whereas paired-end sequencing also creates sequences the read in the opposite direction. The difference behind the two approaches hinges on throughput, whereas the production and availability of single-end tags greatly exceeds the number of complete cDNA fragments nearly ten-fold. Another important advantage of sequenceing-based methods for TSS identification, is they permit one to measure transcript abundance, which allows for a more quantitative analysis of gene expression levels than that of previous methods (Sandelin et al., 2007).

Among these methods, is Chromatin immunoprecipitation (ChIP) (Nelson et al., 2006), which underlies an approach centered around the chromatin immunoprecipitation of promoter-associated protein, and thus provides a genome-wide approach to promoter identification (Nelson et al., 2006). ChIP involves the use of antibodies to target DNA-bound factors commonly associated with promoter activity. The DNA-bound proteins are

fixed via cross-linking, and the genome is fragmented. The DNA-bound proteins are extracted using immunoprecipitation and then isolated. In order to characterize the genomic location of the DNA fragments--and thus the cis-regulatory region—the fragmented reads are sequenced then aligned to the genome.

TSS profiling methods

TSS profiling methods are the most widely used, and arguably garner the most success in terms of methods available for promoter identification. Among the first of the high-throughput, genome-wide sequence-based methods to emerge was the 5`-end Serial Analysis of Gene Expression, or SAGE (Hashimoto, et.al, 2004). This method was useful for the identification of the 5`-end of transcripts, and requires the cDNA library construction of oligo-capped mRNA preparations. The protocol centers around the enzymatic replacement of the mRNA cap structure with either a restriction endonuclease site or a restriction enzyme site. The cleaved transcript is reverse transcribed to create a cDNA library through the use of a random adapter-primer; the second strand is synthesized using a biotin-bound 5` (forward) primer. The cDNA is cleaved 20 base pairs downstream of the introduced recognition site, and the two fragments with different cap signatures are ligated together to create a hybrid concatemer.

The most widely used TSS profiling method is the Cap Analysis of Gene Expression (CAGE) (Hazuki Takahashi, 2014). CAGE identifies TSSs at large-scale and at single resolution and also provides a measure of promoter activity and gene expression. The CAGE protocol begins by synthesizing cDNA strands from extracted total RNA. A cap trapping method selects the 5`-end of capped mRNAs to associate a biotinylated

linker coupled with endonuclease recognition sites. Following second-strand cDNA synthesis, cDNA molecules are cleaved to separate the CAGE tags from the original input RNA. Amplification and sequencing linkers are attached to the 3`-end of the resulting molecule to facilitate library sequencing. The double-stranded DNA molecules are digested using Mmel, and then ligated and amplified using PCR amplification. Over the years, CAGE has undergone multiple revisions, including nanoCAGE and SLiC-CAGE (Cumbie et al., 2015; Cvetesic et al., 2018) . In general, the major drawback of CAGE methods is that they are limited by the large quantity of total RNA required (5µg of total RNA) and their time-consuming nature (Shiraki et al., 2003).

Another TSS profiling method is known as RAMPAGE (RNA Annotation and Mapping of Promoters and Analysis of Gene Expression; (Batut & Gingeras, 2013a) RAMPAGE selects 5`-complete molecules, allowing for promoter activity profiling using paired-end cDNAs. RAMPAGE incorporates two important advances. First, it utilizes the use of template-switching reverse transcription (TSRT), which naturally introduces 3-4 untemplated C's, to facilitate the addition of adapter sequences to the ends of the cDNAs (Schmidt & Mueller, 1999). Second, it uses cap-trapping (Takahashi et al., 2012), introduces the biotinylation of the capped RNA modules which aids the downstream pulldown of capped transcripts and their complementary cDNAs (Batut & Gingeras, 2013b). The RAMPAGE protocol consists of reverse transcription of an RNA sample, followed by cap oxidation and biotinylation. After RNase I digestion, the first-strand RNA-cDNA complex is pulled down via streptavidin-coated beads, and the second round of PCR amplification takes place, creating double-stranded cDNA libraries.

Subsequent size selection, which removes oligionucleotide complexes, results in a completed RAMPAGE library (Batut & Gingeras, 2013b)

STRIPE-seq. Previous methods of TSS profiling have laid the foundation for advancement and subsequent refinement of the mapping of transcription start sites; however, these approaches tend to be laborious, time-consuming and expensive, creating a barrier to entry for laboratories wishing to perform TSS profiling in their own laboratories. However, the recent development of STRIPE-seq (Survey of Transcription Initiation and Promoter Elements; Policastro et al, 2020), has made TSS profiling much easier. STRIPE-seq is a cost effective, and straightforward protocol that can be completed in any well-appointed molecular biology laboratory.

Like RAMPAGE, STRIPE-seq relies on template-switching reverse transcription (TSRT) to capture 5'-complete mRNAs. A key aspect of the STRIPE-seq protocol is the biotin modification, which alleviates the influx of oligio complexes, which have been a signigifcate barrier in TSRT reactions without this addition. During the TSRT, the reverse transcription oligonucleotide (RTO) and the template switching oligonucleotide (TSO) bind to the transcripts provided by the input total RNA via reverse transcriptase's introduction of 3 untemplated C's (see Figure 2). The RTO contains a 5 base-pair random primer and a barcode adapter to facilitate downstream sequencing, while the TSO contains an 8-nt unique molecular identifier (UMI), which facilitates the removal of PCR duplicates, a TATA spacer to prevent TSO invasion (Tang et al., 2012), and a biotin cap to alleviate the potential for the formation of TSO concatemers. Following TSRT, a bead-based size selection is conducted, which removes any primer dimer formations. Next, the samples are amplified during PCR amplification and a second, more intensive solid-phase

reversible immobilization using bead-based size selection to remove any remaining primer dimers, leading to the final STRIPE-seq library prior to sequencing. The STRIPE-seq protocol is illustrated in detail in Figure 4.

Promoter Architecture

The architectural properties of a promoter have proven to be important insights into a promoter's functionality and therefore gene expression. Promoter architecture refers to the distribution of the TSSs within the genomic region in which transcription initiation occurs. In eukaryotes, transcription initiation commences when the RNA polymerase II complex associates at the core promoter region and begins to synthesize the first nucleotide, known as the transcription start site (TSS). However, rather than commencing at a single, pre-determined nucleotide each time a gene is transcribed, initiation commonly occurs at an array of genomic positions located within the core promoter region, spanning around 75bp (-50 to +25). Using this model, it is beneficial to define a TSS as a transcription start region (TSR), as it accounts for the distributed nature of start sites at that have been identified in the genomes across eukaryotes, including *Drosophila melanogaster* (Rach et al., 2009), *C. elegans* (Saito et al., 2013) and mammals (Carninci et al., 2006)(FANTOM Consortium et al., 2014). When investigating the properties of TSRs, it was noted that they have two major shape classes, peaked and broad (Rach et al., 2009). Hoskins and colleagues (Hoskins et al., 2011) characterized promoter shape quantitatively, using a modified form of Shannon entropy (Shannon, 1948), which is discussed in further detail below, in conjunction with the shape index.

In some instances, TSSs span across a wide range of nucleotides with roughly equal promoter usage across each individual start site. TSSs that resemble this architecture are referred to as “broad” promoters, owing to their distinctive shape. On the other hand, TSS distributions that can be characterized by the possession of a single major TSSs have been described as a “peaked” or “sharp” promoter (Figure 1). Several studies, employing TSS profiling to investigate promoter architecture in a range of model organisms, have revealed that peaked promoters commonly regulate tissue-specific genes, whereas broad promoters are associated with housekeeping genes (Rach et al., 2009; Hoskins et al., 2011).

Regulatory Elements of Metazoan Promoters. In eukaryotes, the core promoter is commonly understood to refer to the region of DNA immediately adjacent to the TSS, usually spanning 50 base-pairs (bp), flanking both sides of the TSS (Haberle and Stark, 2018). The core promoter is to act as a binding platform for the transcriptional machinery, and to position the transcription initiation complex. The transcription initiation complex contains GTF and DNA-dependent RNA polymerase, and is responsible for the construction of mRNA molecules via synthesis of the DNA template (Haberle and Stark, 2012).

Analysis of the core promoter and its sequence motifs has revealed the evolutionary conservation of several interchangeable sequence elements localized near the TSS, including the Inr and the DPE element (Juven-Gershon, Hsu, & Kadonaga, 2008)(Juven-Gershon, Hsu, Theisen, et al., 2008). Functionally, these motifs are responsible for the association of the pre-initiation complex, or PIC, which houses the

GTFs and RNA polymerase II prior to binding (Ponjavic et al., 2006). Despite the core promoter's crucial responsibilities, the effects of gene expression, as well as the conservation of the PIC across the metazoan tree of life, the core promoter has evolved a plethora of sequence features illustrating an aspect of flexibility.

Through the divergence and shuffling of core promoter motifs, the core promoter displays an intriguing pattern of motif composition, including lineage-specific novelties such as DPE, which is conserved across the metazoan lineage (Sandelin et al., 2007). Furthermore, the area surrounding the core promoter region contains several characteristic motif elements, including the TATA box, BRE motifs, MTE, DTE, and DCE1-3, some of which segregate according to promoter classification (Haberle & Stark, 2018)(Juven-Gershon et al., 2008). The core promoter elements do not represent a universal attribute of promoter sequence architecture, but rather a dynamic inclusion in which some promoters lack all core promoter motifs (Juven-Gershon et al, 2008). The functional attributes, locations, and phylogenetic distribution of core promoter motifs are discussed below (Figure 2).

Core Promoter Motifs

The TATA Box. A subset of regulatory regions contain a functionally relevant TATA box motif (present in 10%-20% of surveyed metazoan regulatory sequences) which was previously thought to be the hallmark of a *cis*-regulatory region, as it was the first core promoter motif to be recognized (Ponjavic et al, 2006) (Goldberg, ML, 1979). The TATA box has a consensus sequence of TATAWAAR in metazoans, giving this element its name. The TATA box is one of the regulatory signatures commonly present in promoters and is responsible for the binding of the TBP subunit of TFIID (Kadonaga, T. James, 2012) (Haberle, Vanja, and Stark, Alexander, 2018), an important GTF. The TATA box is found in a precise location relative to the TSS, and is located at -31 or -30, i.e. 30 or 31bp upstream of the Initiator motif (Inr) (Kadonaga, 2012). The location of the TATA box is tightly restricted due to the precision required in the binding of the PIC at the initiator site; this distance-based relationship with the initiator sequence (Inr) and the TATA box represents the functional determinant of TSS selection (Ponjavic et al., 2006).

The Initiator Motif. The Initiator (Inr) sequence is a more customary transcriptional element when compared to the TATA box as it is more common within the genome, though its consensus sequence is somewhat flexible. For example its consensus differs, between *Drosophila* (TCAKTY) and humans (YYANWYY). The Inr motif serves an important role in initiation as it aids in the association of extraneous TFFIID components due to its overlapping proximity to the TSS [+1] (Haberle and Stark, 2018). More specifically, basal transcription initiation hinges on the association of TFIID to the Inr element (Kadonaga, 2012).

The B Recognition Element Motif. The B Recognition Element motif, or BRE, is a promoter element that acts in conjunction with the TATA box to regulate basal transcription activity, through both the activation and repression of initiation (Kadonaga, 2012). Additionally, the presence of two BRE motifs has been noted, including BREu (BRE upstream) and BREd (BRE downstream), although their consensus sequence, SSRCGCC, and evolutionary presence ranging from Archaea to human lineages, has been evolutionarily conserved (Lenhard et al, 2012; Kadonaga, 2012). Alongside the TATA box, BRE represents the most ancient of the core promoter elements (Juven-Gershon et al, 2008).

The Downstream Core Promoter Element and Motif Ten Element. In terms of the downstream regulatory motifs, which are still located within the confines of the core promoter, is the downstream core promoter element (DPE), which was first isolated in TATA-less promoters (Burke et al., 1996) and contains a consensus sequenced of RGWYVT in *D. melanogaster* (Juven-Gershon et al., 2008). In such promoters, the DPE is accompanied by the Inr element, which facilitates proper Inr-DPE spacing, allowing for the aggregation of TFIID subunits (Haberle and Stark, 2018). The distribution of DPE is known to be 29 to 31bp downstream of the TSS. In *Drosophila*, adjacent to the DPE, is the motif 10 element (MTE), which occupies base pairs +19 to +28. Both of these elements have been linked to four downstream regions, the first and second are required for the MTE, and the second and third are linked to the function of the DPE (Lenhart, 2012). Both the DPE and the MTE represent preserved core promoter elements across the metazoan lineage (Juven-Gershon et al., 2008).

The Downstream Core Element Motif. The DCE motif (Downstream Core Element), represents another downstream sequence feature with distinctive architectural properties, and is linked to the presence of a TATA box motif. Comprised within the DCE motif are three significant sub-elements which occur between +6 to +34bp (Juven-Gershon et al, 2008); the first of these elements is the sequence CTTC, which occurs from +6 to +11, CTGT occurs from +16 to +21, and AGC occurs from +30 to +34 (Juven-Gershon et al., 2008).

The Significance of Promoter Shape Classes. The presence of *cis*-regulatory diversity originally was described through the classification of two distinct models of promoter architecture, high-CG and low-CG (Carninci et al., 2006). Not only did the two promoter classes exhibit differing nucleotide compositions, they also displayed divergence through the distribution and abundance of TSS tags. That is, high-GC promoters signified a broad distribution of TSSs, an overlap with and CpG island, and commonly regulated widely expressed, or developmentally crucial genes (Deaton et al, 2011). The other side of the spectrum is represented by a focused, or peaked TSR, which initiates at a single nucleotide position; these are associated with low-GC sequences. The peaked promoter class associated with tissue-specific gene expression (Lenhart et al, 2012) (Figure 1).

In an investigation of core promoters conducted by FitzGerald and colleagues (FitzGerald et al., 2006), *Drosophila* promoters were grouped into three classes, type I, type II, and type III (Lenhard et al., 2012); the first of these classes included tissue-specific promoters with an associated TATA box and Inr motif (Lenhard et al., 2012). The second promoter class were those containing a DRE element or a non-universal

selection of novel motifs; together these encompass constitutively expressed, "housekeeping" genes. Lastly, promoters containing an Inr element or a combination of an Inr element as a DPE were grouped into the third class. The third class of promoters are associated with developmentally linked genes, which require synchronized and precise expression (FitzGerald et al., 2006; Lenhard et al, 2012).

FitzGerald's investigation also provided insights into the classification of mammalian promoter architecture and functionality. Mammalian promoter architecture differs from the *Drosophila* model due to the absence of TATA boxes and overlap with CpG islands; however, a framework for classification was soon developed that resembles the version implemented in *Drosophila*. Type I promoters are characterized by promoters containing a TATA box and low CpG occurrence within vertebrates; they are usually associated with the regulation of tissue-specific genes. Housekeeping genes, which are referred to as type II promoters, commonly contain an overlapping CpG island and are TATA-depleted within mammals. Lastly, Type III promoters are associated with developmental genes and usually encompass several overlapping CpG islands extending into the gene (Lenhard et al, 2012). The accumulation of this knowledge can aid in the analysis of the conservation of the promoter motif elements and their presence across populations. It also dovetails an analysis into the types of elements segregating according to gene specific patterns, which can be investigation between clones, between populations, and between species.

Further implicating the importance of promoter shape are investigations which identify promoter shape as a molecular trait that evolves independently and varies across populations of inbred *Drosophila* lines. In their investigation, Schor and colleagues

suggest that promoter shape has important implications in the alteration of expression noise and subsequent evolution as natural genetic variants affecting the shape of a promoter cause an increase in the expression noise. (Schor et al., 2017). Furthermore, these investigations which utilize natural variants affecting TSS usage uncovered evidence of adaptive selection within broad promoters further illustrating the occurrence of promoter evolution (Schor et al., 2017).

Daphnia as a Model Organism

Daphnia is a micro-crustacean within the metazoan lineage, and the first crustean to have its genome sequenced (Colbourne et al., 2011). *Daphnia* are commonly referred to as a "water flea" due to their sporadic, jumpy movements through the water, facilitated by the flapping of their large antennae, which act more in resemblance of arms than actual antennae. In a normal growth state in which there is an abundance of resources, *Daphnia* reproduce asexually, through the development of diploid eggs directly in the anterior brood chamber of the female *Daphnia* (Stollewerk, 2010). These eggs then take about 3 days to be released into the aquatic environment, where they develop through six larval stages before reaching adulthood. In times of food scarcity or overcrowding, the female Daphnia can reproduce sexually, through meiosis, and generate a resting egg, which can endure many varying conditions during its dormancy stage, before being triggered into further stages of development by an external stimuli (such as rising temperatures) (Stollewerk, 2010). There are some reported Daphnia lines that cannot reproduce sexually, and also are unable to contribute to the generation of male *Daphnia* (Ye et al., 2019). Additionally, culturing of the *Daphnia* is a relatively easy task, and

plethora of populations can be maintained using minimal resources and space; the lab currently maintains a library of *Daphnia* stocks from which various clones can be sequestered and produced. Previous TSS profiling has demonstrated the usability of *Daphnia pulex* for this type of analysis, namely through the previously generated *Daphnia* Promoter Atlas (Raborn et al., 2016), which provides a catalogue of active promoters across three developmental stages (sexual females, asexual females, and sexual males). *Daphia pulex* is also represented by its high levels of heterozygosity, large effective population sizes, and transparent appearance, further suggesting its plausibility as a pivotal model organism and system of study (Haag et al., 2009; Tucker et al., 2013). Previous investigations have revealed the presence of eight core promoter elements, including TATA and Initiator (Inr) (Raborn et al., 2016). Additionally, an average of just over 12,000 TSRs (n=12662) was reported by this analysis, as well as the presence of two segregating promoter shapes, as suggested by investigations in other organisms. Broad promoters were determined to have higher transcriptional activity (Raborn et al., 2016), consistent with what was reported in *D. melanogaster* (Hoskins et al., 2011).

Using the foundation built by previous investigations into TSS profiling, I sought to conduct an in-depth analysis of *cis*-regulatory evolution at both the population and the species level, using naturally occurring *Daphnia* populations and *Daphnia* species, characterized by a span of evolutionary time. *Daphnia*, a major ecological and emerging genetic model organism, was selected for this analysis based on several compelling factors. First *Daphnia* represents the first microcrustacean with a sequenced and annotated genome, including the various assemblies for various species (*D. pulex*, *D. pulicaria*, and *D. obtusa*), therefore providing a basis for mapping transcripts.

Moreover, *Daphnia* have the ability to reproduce both asexually, through parthenogenesis, and sexually through the formation of a resting egg (Innes, 1997). Because sexual reproduction is exploited during times of distress and/or scarce resources, asexual reproduction can be supported in a lab environment through daily feedings, segregation into additional habitats when crowding occurs, and fluid replacement (Innes, 1997). Because of this feature, a single *Daphnia* can produce many clonal replicates through parthenogenesis.

The proposed species analysis will be conducted using TSS profiling data, generated by STRIPE-seq originating with two other species, *D. pulicaria* and *D. obtusa*. *D. pulicaria* represent a recently diverged lineage from *D. pulex*, which is morphologically indistinguishable to the naked eye. Phylogenetic estimates suggest that *D. pulicaria* diverged from *D. pulex* 200,000 years ago. The second species, *D. obtusa*, represents an outgroup to *D. pulex*., with a divergence time of ~2 million years ago. Taken together, these species provide a ranging scope of evolutionary divergence time, and therefore allow for a robust interpretation of *cis*-regulatory divergence across differing time points.

Research in *D. pulex* is aided by an abundance of well-characterized naturally occurring populations, found in freshwater ponds with varying degrees of permanence (Lynch, 1983). The populations utilized in this analysis are dispersed throughout the Midwestern US and Canada, Oregon, and Arizona. We sought to use the dispersed populations as a basis for investigation to study TSS evolution in the context of population genetics, through the identification of the promoter regions, and subsequent investigation into the promoter architecture (Figure 3 and Table 1).

Overall, this analysis provides the methodology for the development of a promoter library data set using a novel TSS profiling method in a new model organism. The generated libraries include promoter identification using clones of the same populations, clones of differing populations, and clones from differing species.

Promoter Evolution

Although a number of TSS profiling studies have been conducted, they have been largely limited to analyses in a handful of major model organisms, namely *S. ceravisae*, *Drosophila*, and *C. elegans*. (Carninci et al., 2006; Hoskins et al., 2011). However, a handful of investigations into the evolution of promoters have been conducted (Li et al., 2018; Schor et al., 2017; Main et al., 2013). These investigations are often constrained by time and financial input required by previous methods (e.g. CAGE), as well as annotation and genome presence, prohibiting wide-spread investigations across populations and species. The investigation conducted by Main and colleagues specifically, attempts to investigate promoter discrepancies across related species of *Dropsophila*, in hopes of determining the evolutionary principles that govern the transcription start sites. Additionally, as is the case with Main's investigation, much of these investigations into promoter evolution fail to generate adequate read coverage and sequencing depth. Nonetheless, TSS evolutionary inquiries have revealed a plethora of important knowledge regarding the development and alteration of *cis*-regulatory regions. Some of these investigations were limited by insufficient promoter coverage. Analyses in these select organisms are also criticized for their expansive evolutionary distance, as illustrated by the presence of unpaired TSS in a select species of *Drosophila* and inability

of alignment across mammals, greatly hindering the conclusions of the study. Main et al.'s investigation of promoter evolution in four *Drosophila* species laid the foundation for further investigations into TSS evolution at the species level, although it fell short in generating in-depth, high coverage libraries, further hindered by the evolutionary distance between the selected species. Summarizing the findings in *Drosophila*, Main and colleagues revealed that the location and activity of promoters is largely conserved across evolutionary distances, although certain species presented with elevated sequence divergence localized upstream of the TSS, suggesting an elevated mutation rate at these sites (Main et al, 2013). The technical limitation of this investigation, including the divergence time between samples, illustrates the necessity of an approach that utilizes populations and species with less vast levels of divergence.

Investigations into the evolutionary characteristics of promoters (Schor et al., 2017), have uncovered many trends that are beginning to emerge; promoter involved in the regulation of housekeeping genes depict the highest level of conservation, including those that regulate the production fibroblasts, chondrocytes, and pre-adipocytes. Conversely, TSSs restricted in their function to a single cell type are much more evolutionarily flexible, and therefore represent patterns of loss and gain (FANTOM Consortium, 2014). In turn, the peaked TSS architecture and the corresponding defining motifs is more common in TSSs that have relocated (Main et al., 2013). Genes which are specific to the expression of T-cells, macrophages, dendritic cells, whole blood, and endothelial cells, fall into the latter category, representing a swiftly progressing immune system. The evolutionary conservation of promoter shape remains intact not only across species and populations, but also throughout developmental stages, in which embryos and

adult regulatory regions are defined by 95% shape similarity in *Drosophila* (Hoskins et al., 2011). That is, promoter shape classification is vastly similar through differing life cycle stages.

Investigations into the emergence of regulatory regions, including those conducted by Main and Li, and their relationship to the generation of diversity is a topic that has garnered attention of late. New promoters commonly emerge via random mutations or as an effect of relaxed selection (Main et al, 2013), a concept further expanded to indict retrotransposons and the proximity to existing transcriptional machinery. Furthermore, the presence of viable sequence properties and motifs (Li et al, 2018) facilitate the emergence of a novel regulatory region, usually within repressed chromatin genomic regions. Following the advent of a novel TSS, subsequent evolution facilitates the advent of alternative promoter usage (Li et al., 2018), which leads to a phenomenon that has been noted in humans and *Drosophila* (Main et al., 2013) as well as suggested in humans and yeast.

At the sequence level, evolution of emerging promoters is characterized by rapid evolution which decreases in intensity and levels off as a stable genomic environment is generated, allowing for the reduction of transposition capacity (Li et al., 2018). Prior to the maturation of a novel TSS to an older TSS and expansion of their limited regulatory role, the new TSS often experiences weaker transcription and is ill-defined in terms of a regulatory agenda or tissue-specificity. Their heightened levels of evolution displayed through limited human studies (Li et al., 2018) aids in the introduction of chromatin accessibility and histone modifications, a characteristic absent in the new TSS, which drives the increase of expression and the maturation of a new promoter.

Taken together, this manuscript outlines the methods development, including the introduction of a novel TSS method and adaptation to a new model organism; the complete composition of the computational and experimental methods, including suggested areas of future direction and analysis are contained within this manuscript. Additionally, the generation of the minimal STRIPE libraries and the computational identification of the promoters was completed in conjunction to this thesis. The methodology is described below.

CHAPTER 2

METHODS

Experimental Methods

***Daphnia* Culturing and Maintenance.** In this study, various *Daphnia* population were isolated from existing stock cultures within the pre-existing inventory (Table1). Initial inoculation involved the identification and subsequent isolation of ~5 adult females currently carrying eggs within their brood chamber. This facilitates the clonal reproduction properties discussed above and limits the quantity of males present at a given time. The cultures were housed within a liquid media containing ~75% autoclave sterilized distilled water, and ~25% cultured lake water. The *Daphnia* were fed a mixture of *Scenedesmus* (~100,000 cells/mL) and autoclaved distilled water about every other day. Water was replaced due to accumulation of waste and excess algae about every 2 weeks for the duration of the study.

Total RNA Isolation. Total RNA Isolation was performed using a Zymo Direct-zol RNA kit (Zymo R2052; Zymo Research Irvine, California) and Trizol reagent (Invitrogen, Waltham, Massachusetts). Adult, female *Daphnia* were visually identified and isolated under a dissecting microscope via morphological differences and/or presence of resting egg or occupation of brood chamber, and 12 individuals were extracted for each sample. The culturing medium was removed from the sample tube and 200mL of TRIzol reagent was immediately added. Whole *Daphnia* individuals were then homogenized using a mortar and pestle using a motor-driven grinder, which was followed by a spin-column-based RNA isolation using the standard Direct-zol protocol (Zymo Research Corp.). Following RNA isolation, the RNA integrity and abundance was

analyzed using the Agilent Tapestation and the Thermo Fisher Qubit, respectively. RNA samples were stored at -80°C until use.

TSS Profiling Using STRIPE-seq. STRIPE-seq (Policastro et al., 2020) was selected as the method for TSS profiling due to its low initial RNA input, minimal financial constraints, and overall time to construct a library. TSS library construction utilized STRIPE-seq to sequence the 5` ends of the RNA, using an input of 200ng of total RNA. Following an optional TEX digestion to remove the ribosomal RNA , STRIPE-seq uses template switching reverse transcription to construct a cDNA library of the total RNA. Reverse transcriptase (I used Superscript II Reverse TranscriptaseTr from Thermo Fisher Scientific) synthesizes 3 untemplated Cs, which facilitates the binding of the primers, the transcription oligionucleotide and the template switching oligionucleotide. These primers contain a 5-nucleotide random sequence, an Illumina TrueSeq P7 barcode adapter, a unique molecular identifier, a TATA spacer, and a biotinylated cap. The primers were purchased from IDT; the sequence of the TSO is 5` CCTACA CGA CGCTCTTCCGATCTN; the sequence of the FLO is 5` AATGATACGGCGAACCACCGAGATCTAC; the sequence of the RLO is 5` CAAGCAGAAGACGGCATACG. A bead-based size selection (using RNA Clean XP) was conducted both post and prior to PCR amplification, resulting in a finalized library to be sequenced using Illumina Next-seq 150 (Center for Genomics and Bioinformatics, Indiana University) .The complete protocol can be found on Protocols.io (https://www.protocols.io/view/stripe-seq-library-construction-bdtri6m6). I also utilized Thermo Fisher Scientific's King Fisher Flex System to automate the size selection bead cleanups. In the first cleanup, following the reverse transcription reaction, I used a ratio

of 0.8 RNA Clean XP beads to product, to size select the library and remove any primer dimers. In the second cleanup, which occurs after the PCR amplification step, I used a 0.6 ratio of RNA Clean XP beads to product, followed by a ratio of 0.7 beads to product to remove primer dimers

Automation of Size Based Bead Selections. Relatively early on in the investigation, after we were able to optimize the RNA extraction protocol and ramp up STRIPE-seq library construction we ran into a bottleneck at the bead-clean step. This step was by far the most time consuming step in the STRIPE-seq protocol, especially when attempting to run multiple samples. Luckily, I was able to automate this process by programming the Kingfisher Flex Magnetic Particle Processer (Thermo Fisher Scientific) to complete the magnetic bead clean-ups. This allowed us to complete up to 96 libraries, in the time that it took to do a single one by hand, speeding up the time to completion of a library drastically.

RNA-seq Analysis. We carried out RNA-seq analysis using populations KAP and LPA to determine if gene product differences occur, and if so, could these differences be attributed to *cis*-regulatory differences. I extracted total RNA from four adult, female replicate samples in each population (KAP and LPA) using the same protocol as described above for the RNA extraction, including the use of the Trizol reagent and homogenization. RNA samples were sequenced using Illumina Next-seq 150 (Center for Genomics and Bioinformatics, Indiana University) and the resulting libraries were analyzed computationally using the edgeR package to complete the differential expression analysis (https://bioconductor.org/packages/release/bioc/html/edgeR.html).

Computational Methods

I utilized the GoSTRIPES Singularity package (https://github.com/BrendelGroup/GoSTRIPES) to prepare and analyze the STRIPE sequences. The first step within this pipeline involves quality control and filtering of the sequenced libraries, which begins with identification of the R1 read through he identification of the UMI, spacer, and 3 G insertion (NNNNNNNNTATAGGG). The TATAGGG following the UMI is trimmed using the fastx toolkit, and rRNA contamination reads are filtered out using Tagdust (Lassmann et al., 2009). Tagdust was also utilized to remove the low-complexity reads. FastQC files are generated for both the raw and processed fastq files, to ensure the construction of quality libraries. Processed reads were then aligned to the latest version of the *D. pulex* genome assembly (PA42 v.4.0; Ye et al., in Preparation) using Burrows-Wheeler Aligner (Li & Durbin, 2009). At this point, unpaired reads and non-primary alignments were removed and UMI were de-duplicated using samtools. Finally, the TSSs were identified, associated with gene annotations, and clustered into TSRs using the package TSRchitect (Raborn et al., 2017) (https://github.com/ssnyde11/tsrchitect_tsrexplorer_figures). Replicates were merged using the Rdata file generated during the TSS identification and analysis, resulting in population samples, rather than single replicates (https://github.com/ssnyde11/tsrchitect_mergedTSRs). TSS and TSR analysis, including count normalization, construction of correlation plots, genomic location analysis, and promoter feature characterization was completed using an adaptation of the TSRexplorer package (https://github.com/rpolicastro/tsrexplorer) (https://github.com/ssnyde11/tsrchitect_tsrexplorer_figures). Subsequent figures, such as

the TSS/TSR histograms, violin plots, and box and whisker plots were generated using original scripts (https://github.com/ssnyde11/TSS_figure_generation)

CHAPTER 3

RESULTS AND DISCUSSION

RNA-seq and Differential Expression Analysis

First, it is crucial to determine if there are gene expression differences present between various populations of *Daphnia*, as this investigation serves as a proof-of-concept; if there are variances in the production of transcripts between the samples, it is possible to determine if these differences are caused by alterations in *cis*-regulatory regions or their use. In order to detect potential expression dissimilarities between the *Daphnia* populations, I conducted an RNA-seq and a subsequent differential expression analysis. Two populations, KAP and LPA, were arbitrarily selected to serve as the model populations during this analysis. The RNA from four, adult female *Daphnia* from each population was extracted (8 total samples). Following RNA sequencing, the quality of the fastqs was validated using fastQC (https://www.bioinformatics.babraham.ac.uk/projects/download.html). Alignment to latest version of the *D. pulex* genome assembly (PA42 v.4.0; Ye et al., in Preparation) was completed using Burrows-Wheeler Aligner (Li & Durbin, 2009). A heatmap and smear plot limited to the known differentially expressed genes was constructed using the Empirical Analysis of Digital Gene Expression Data in R package (edgeR; Robinson et al., 2010). The results indicated consistency between replicates and identified 1,014 differentially expressed genes out of the approximately 20,000 (p=.05) known within the *Daphnia* genome. Although it is difficult to isolate extraneous conditions, such as differences in the amount of food available, or the location within the culturing room, the results of this investigation indicate that there are quantifiable gene expression

differences present between different populations of *Daphnia*. This facilitates downstream analysis into the identification and analysis of transcription start sites, in order to determine if these differences are due to alterations in these regulatory regions or their utilization (Figures 5 and 6).

Transcription Start Site Profiling

Read Processing and Alignment. Following Illumina sequencing, the resulting libraries, including BRV, W-17, LPA, LPB, NFL, OA15, OA85, PA42, POV12, POV84, TEX36, and LHM were identified for downstream analysis. Libraries with small file sizes (less than 1MB) were discarded, along with those with an aligned read count less than ~100,000; these libraries included 3 BRV samples, 4 LHM samples, a POV84, a POV12, an LPA replicate, a TEX36 replicate, an OA85 replicate, and a PA42 replicate. The distribution and quantification of raw reads, processed reads, and TSSs is displayed in Table 1. The average number of raw reads from sequenced libraries (fastqs) is 21,156,324 across eight populations, including BRV, LPB, OA15, OA85, POV12, TEX36, NFL, and POV84. The average number of processed and aligned reads across all populations was 648,883, meaning that an average of 3% raw reads mapped to the genome. This is likely due to several factors, including algae contamination, the presence of rRNA, and the occurrence of low quality, or unpaired reads that were removed during processing and quality control; the optional TEX digestion within the STRIPE-seq protocol could be utilized to improve this statistic through the experimental digestion of ribosomal RNA. Finally, the average number of identified TSSs across all eight populations was 206,396, contributing to the identification of an average of 14,471 promoter regions (Table 2).

Replicate Validation

Another aspect of quality control that was completed was the analysis of the similarity between replicates. For this analysis, I defined a replicate as the resulting STRIPE library of 12 individual *Daphnia*, cultured from the same stock *Daphnia* (stocks represents lab-wide inventory different population or species cultures). Due to the asexual nature of *Daphnia* when housed in ideal conditions, the offspring should consist of clonal replicates, indicating a high level of genetic inter-replicate similarity. Ensuring conservation between the replicates safeguards against accidental contamination or improper lab handling, which could result in skewed results, although environmental factors which could be affecting gene expression were not controlled for. To validate the similarity between the samples, TSS correlation plots were generated using the R package TSRexplorer. (https://github.com/rpolicastro/tsrexplorer). TEX36 was selected as the representative population, although correlation plots were generated for all samples. The resulting correlation plot (Figure 7) yielded an r^2 value of .919, between replicate 1 and 2, indicating a strong correlation between the TEX replicates. The graph on the left is a plot of the overlapping TSSs between the TEX replicates. The clustering of the dots represents the overlap in the expression level at the overlapping TSSs, therefore dots which are closer to the y=x line are overlapping TSSs with overlapping expression levels at those sites.

TSS Quantification and TSR Clustering

The general premise behind the identification of TSRs using the TSRchitect package involves utilizing the genomic location of the mapped reads (now called tags) to identify TSSs. Essentially, the resulting TSSs are sites at which a certain number of tags is reached. During the analysis of transcription initiation using TSRchitect, the identified TSSs neighboring TSSs, were clustered together to create a TSR. This model indicates a need for a numeric value that indicates occurrences of transcription initiation, but is not affected by artifacts or spurious transcripts; this value is defined as the TSS threshold. The TSS threshold is the number of required tags per site to indicate an area of transcriptional activity above the noise level. Because this figure can be skewed due to the differing numbers of reads, I identified the TSSs multiple times using the TSRchitect pipeline, each with differing TSS threshold values, spanning from 2 to 8 tags. This range was selected because the current literature suggests that the majority of TSSs are marked by a small number of tags, usually between 2 and 5 (Balwierz et al., 2009); furthermore, a small number of aligned reads at a given position were demonstrated to be *bona fide* TSSs, as indicated in the STRIPE-seq documentation (Policastro et al., 2020). This procedure was completed on both the individual replicates, and the merged TSS files (Figure 8 and Figure 9). The resulting TSR counts represent the number of identified TSRs, or, in other words, the number of clustered tags meeting or exceeding the annotated TSS threshold.

Next, the TSSs must be clustered into TSRs in order to represent the current mode of thinking surrounding the distribution of transcription initiation. The current

literature suggests that 70% of TSSs have their nearest TSS neighbor within 10-40 nucleotides (Raborn et al., 2017). TSRchitect contains a protocol that associates TSSs located within annotated promoter regions, and TSSs located outside of known regulatory regions and generates histograms for each TSS type at a clustering distance of 20, 30, and 40 (Raborn et al., 2017), and allows the user to choose the appropriate value; the suggested clustering distance is 30, which was utilized in this analysis (Table 3).

Threshold Identification

In order to determine an acceptable TSS threshold for each sample, the threshold of interest was divided by the total number of aligned, processed reads, expressed in millions. For the merged samples, the summation of the read count across all replicates was utilized. This method allows for the normalization of the data, thus ensuring that a population's counts are not skewed due to differences in read counts. The threshold that produces a normalized reads value closest to five was accepted as a suitable threshold. Five was chosen as the value because the majority of TSS profiling data suggests that 2-5 tags captures a real TSS (Policastro et al., 2020), thus filtering out noise without discounting the low tag requirements. The majority (4) of the populations called for a TSS threshold of 3, one was characterized by a threshold of 2, and the other three suggested the use of a TSS threshold of 5. Downstream analysis was then conducted using the file generated by the indicated run of TSRchitect. That is, the TSS file generated through TSRchitect using the acceptable threshold was selected for downstream analysis.

TSRs were identified within each replicate and merged sample, using a different TSS threshold (2, 3, 5, or 8) in order to determine the acceptable threshold given the number of reads. The threshold was divided by the total number of processed, aligned reads for each sample (merged samples were represented by the summation of the reads within all of the replicates). A resulting value closest to five was selected as indicative of the appropriate TSS threshold for each sample. The results of this phase of analysis are displayed in Table 4. The TSS threshold is defined as the number of tags required to differentiate transcriptional activity from artifacts; the resulting number of TSRs represents the number of clustered TSSs which meet the required threshold. The majority of populations were best suited with a threshold of 3, the others diverged towards a threshold of 5, and one sample required a threshold of 2. The process alleviates the possibility of samples with reads counts skewed away from the average to be mis-represented by a lower (or higher) number of TSSs (and therefore TSRs) (Table 4).

Promoter Architecture and Characteristic Findings

Tags and Identified Transcription Start Sites. Following threshold identification and read normalization, I explored number of promoter elements and characteristics within the merged population samples. First, I quantified the number of tags and identified TSSs, which facilitated further exploration into the properties of the promoters within different populations. Across all populations, there was an average of 249,369 tags identified, which, when coupled with the TSS threshold, identified an average of 220,257 TSSs, contributing to an average of 1.13 tags per TSS. When these qualifying TSSs which met the tag threshold are clustered into TSRs using a clustering

distance of 30 nucleotides, an average of 14,471 TSRs or promoters were identified. These figures suggest that there is an average of 15 TSSs per TSR. BRV presented with the lowest number of tags (n=39795) , TSSs (n=34330) , and TSRs (6874), likely due to the low read count of BRV replicate 1 (n=5760382). TEX36 was characterized by the highest number of tags (n= 510045), but a mid-range number of TSSs (n=302,246) and TSRs (17,947), suggesting the presence of an abundance of tags at each identified TSS. Sites with a lower number of tags per TSS tend to correspond to genes with very low expression levels, or "background transcription" whereas populations with a higher number of tags per TSS are likely experiencing an higher levels of tissue-specific transcription (Balwierz et al., 2009). The OAs (OA15 and OA85) presented with a similar number of tags and an average number of tags per TSS that differs from the average by .14 and an average number of TSSs per TSR that differs by from the average by 2.82, which serves as an indication of the extent inter-clone differences, although OA15 clustered into a higher number of TSRs. The quantities discussed about are displayed in Table 5.

As discussed above, the number of tags directly contribute to the number of TSRs present within a sample through the limitation in the number of tags required to form a TSS, imposed via the TSS threshold. Interesting, the distribution of the tags for the BRV, LPB, and NFL populations depicts a clustering around 2, whereas the remaining populations, namely OA15, OA85, POV12, POV84, and TEX36, begin their distribution of tags closer to 3. The former group also is characterized by a congregation of TSSs with a lower number of tags; a more uniform, "tapering" off of the distribution as the plot nears 5 tags (Figure 13).

An important area which is crucial for future investigations using the data set is an analysis into the inter-clone differences to determine if the resulting libraries are similar within the clones. For example, an analysis into the resulting identified TSSs within the two POV samples (POV84 and POV12) and their level of similarity could provide important insights into the conservation of promoters, the validity of the TSS profiling method, and the degree of genomic similarity between two clones within the same population. Additionally, another avenue requiring future investigation is the identification of gene-specific patterns across clones.

Genomic Localization of TSSs. The genomic location of the identified TSSs, can provide important insights into both the validity of the TSS profiling analysis, and also the quality of the current annotation. The majority of the TSSs identified in this investigation are localized in proximal to an annotated promoter region, with a cut off of 250 bp (Figure 11). Additionally, the majority of TSSs investigated in this study associate with intronic or promoter regions, but about a third of TSSs are present in intergenic or downstream regions, further illustrating the complexity of transcription initiation (Figure 12). The results from this analysis indicate a the necessity of a more robust investigation, in conjunction with the annotation, to determine if the presence of artifacts exist within the dataset or the annotation.

Shape Properties of Identified Promoters

TSR Width. As discussed previously, the shape of a TSR has important implications associated with promoter functionality. Promoters, including those in *D. pulex* (Raborn et al., 2016), can be classified into two major shape classes, broad and

peaked, and are each associated with gene function. Given this, an abundance of broad promoters indicates heightened expression of constitutively expressed genes and *vice versa*, because broad promoters commonly regulate housekeeping genes. The TSR width is comprised of the absolute value of the first and last coordinate points of the TSSs that constitute a TSR. The distribution of the TSR width suggests the majority of TSRs remain under 100bps (the highest TSR width present within the first 75 identified TSRs within BRV was 59 and 89 within TEX36). The distribution of the TSR width of the various populations is displayed in Figure 14 and 15.

Promoter Shape. To better understand the TSR shapes within my samples, I explored promoter shapes using the Shape Index (SI), as addressed in Hoskins, 2011. The SI quantifies the tag locations and heights out of the total possibly locations within the promoter region. An SI greater than -1 is classified as "peaked" whereas an SI value less than or equal to -1 is classified as broad, as presented by Hoskins (Hoskins et al., 2011).

$$SI = 2 + \sum_i^L p_i log_2 p_i \quad (1)$$

The SI, which is based on Shannon Entropy, quantifies the shape of the TSR and ranges from completed peaked (SI=2) to broad (negative SI values). Therefore, the SI is anti-correlated with TSR width. (Hoskins et al., 2011). The population samples congregate around SI values of 2 and 1, suggesting that the majority of transcription initiation is occurring a peaked promoters. As illustrated above, sharp or peaked promoters are correlated with the regulation of tissue-specific genes, suggesting that the

majority of transcripts being synthesized are targeted at tissue-specific expression (Figure 16).

Related to the Shape Index statistic is the modified shape index (mSI), which rescales the SI value between 0 and 1 (Raborn et al., 2017). For example, an mSI value of 0 refers to an extremely broad distribution of tags within the promoter region, whereas a value of 1 is assigned to a completely peaked TSR. This metric further solidifies the presence of two isolated promoter shape classes (Figure 17).

CHAPTER 4

FURTHER ANALYSES AND FUTURE DIRECTIONS

Following the construction of a genome-wide TSS map, including the transcription start sites of the various *Daphnia* populations, several additional analyses can be conducted in order to portray the intricacies of *cis*-regulatory diversity. The first topic of interest surrounding the generated data set is harmonizing it with the annotation files, in hopes of improving both datasets. Because TSRchitect associates the TSS profiling data with gene annotations, coupling the two data sets can strengthen the accuracy of both, aid in the identification of mis-annotations, and recognize instances of alternative transcript isoforms. Another immediate analysis that could be conducted would be to determine the degree of similarity between the populations through the characterization of alterations in sequence architecture or analyzing the degree to which promoter shape/class is altered through evolutionary time. Because the RNA-seq analysis depicted gene expression differences present in identically cultured populations, it is important to determine how wide-spread these disparities are. Variations at the nucleotide level within these promoter regions could be driving diversification through the alteration of the mechanisms of gene expression.

On a gene-by-gene or genome-wide basis, the sequence and architectural properties of the promoter regions could be analyzed from a wide and narrow evolutionary lens, such as was proposed for the species analysis. By utilizing the methods outlined in this investigation, in addition to the sequenced *D. pulicaria* STRIPE libraries, a robust interpretation of promoter evolution at the species level could be coupled with the population genomic data presented here. Through the addition of TSS profiling data

from *D. obtusa* (LHM) and *D.pulex* (PA42), a fitting range of evolutionary divergence time can be introduced. *D.obtusa* is upwards of 2 million years diverged from *D. pulex* whereas *D. pulicaria* is 200,000 years diverged, which establishes a scope of evolutionary divergence in which the evolutionary effects on promoters can be analyzed. Similarities and dissimilarities between promoter sequence and class could be identified at each increment of divergence, and the manner in which these regulatory regions are being adapted could be quantified. Additionally, the BRV population represents a pulex-pulicaria hybrid, which could serve as a branching point when examining the evolution between the three *Daphnia* species. Furthermore, the differences in gene expression could also be analyzed, given the evidence presented in this manuscript that the majority of transcripts originate from tissue-specific genes. Using this, genes of interest could be analyzed at each evolutionary level, to determine if certain regions of DNA are becoming more (or less) widely expressed, as evident through alterations in their shape class. Because the shape index offers a quantitative approach to promoter architecture, changes in this architecture, and the effects that these changes have on transcript abundance could also be quantified.

Furthermore, the presence (or lack) of core promoter motifs could be analyzed to a deeper extent. Although certain motifs and elements often cluster together and associate with certain gene classifications, the reasons why certain motifs (or combinations of motifs) may be included or excluded (aside from their bare function) could be uncovered; motifs that are not native to certain organisms could be genetically manipulated and the relative effects observed, non-traditional motif elements could be combined to determine the synergy, if any, that exists; core elements that commonly associate with certain gene

classifications could be cloned into the regulatory networks of differing genes. Although the relative ages of the various core promoter motifs have been described, these elements could be analyzed from an evolutionary perspective to determine the relative relationships between the tree of life and the various core promoter motifs in hopes of determining how and why certain motifs segregate to the organisms and genetic regions that they do.

Finally, STRIPE-seq has shown promise in terms of the ability to identify enhancers as well, although limited knowledge is available about the presence of enhancers and their distribution throughout the evolutionary tree. Nonetheless, STRIPE-seq could usher in a new era in which genome-wide enhancer atlases can be constructed, and the evolutionary relationships that may exist between populations and species can be analyzed through a different lens.

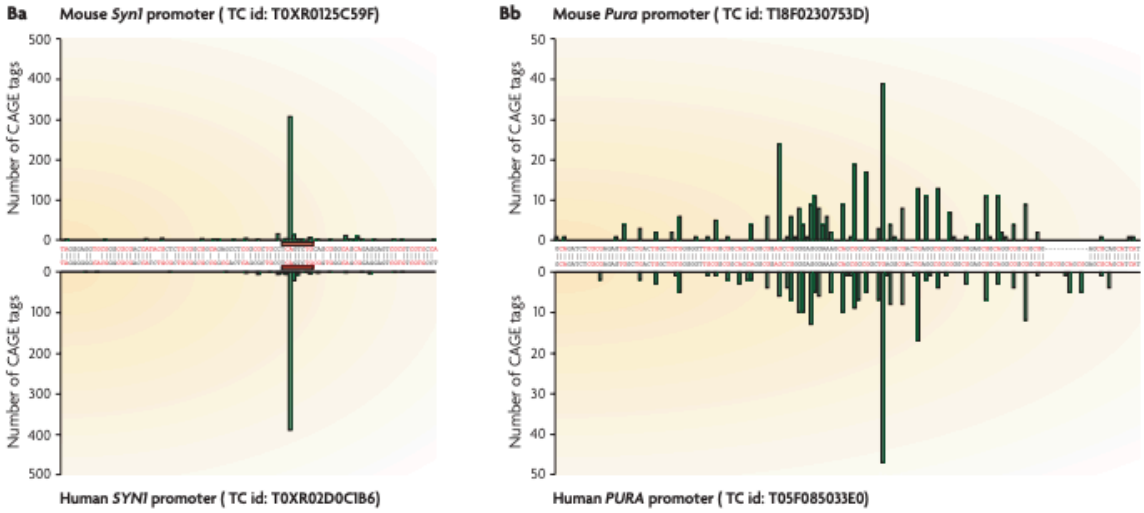

Figure 1. Peaked versus Broad Promoter Architecture [Adapted from (Sandelin et al., 2007)]. Depiction of the general distribution present at a peaked (left) and a broad (right) promoter as seen in human samples. A peaked promoter is represented by a single (or low number) of high-count tags localized at a small genomic locus whereas a broad promoter is defined by a wider distribution of tags, in which there is relatively the same number of tags.

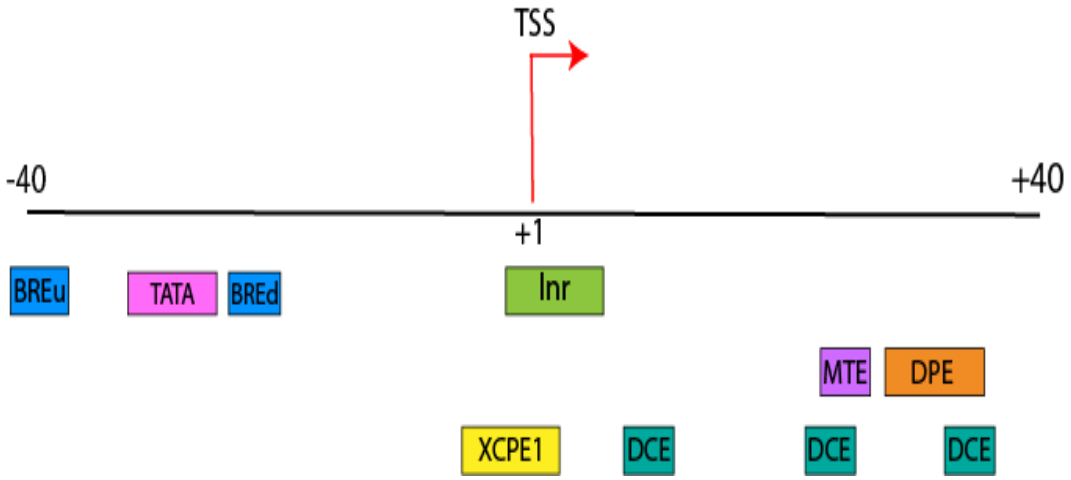

Figure 2. RNA Polymerase II Core Promoter Elements. The transcription start site (TSS) [+1] is flanked by potential promoter elements that are typically found at characteristic positions within the core promoter (shown here as [-40 to +40]). Not every promoter is defined by these motifs, but rather a combination of various elements, although there are instances of transcription initation occurring at genomic positions in which none of the elements are present. It is useful to note that there is not a standard set of elements; rather, a combination of the elements is observed, and is can be associated with the type of gene that it regulates[Adapted from Juven-Gershon, Hsu, Theisen, et al., 2008]).

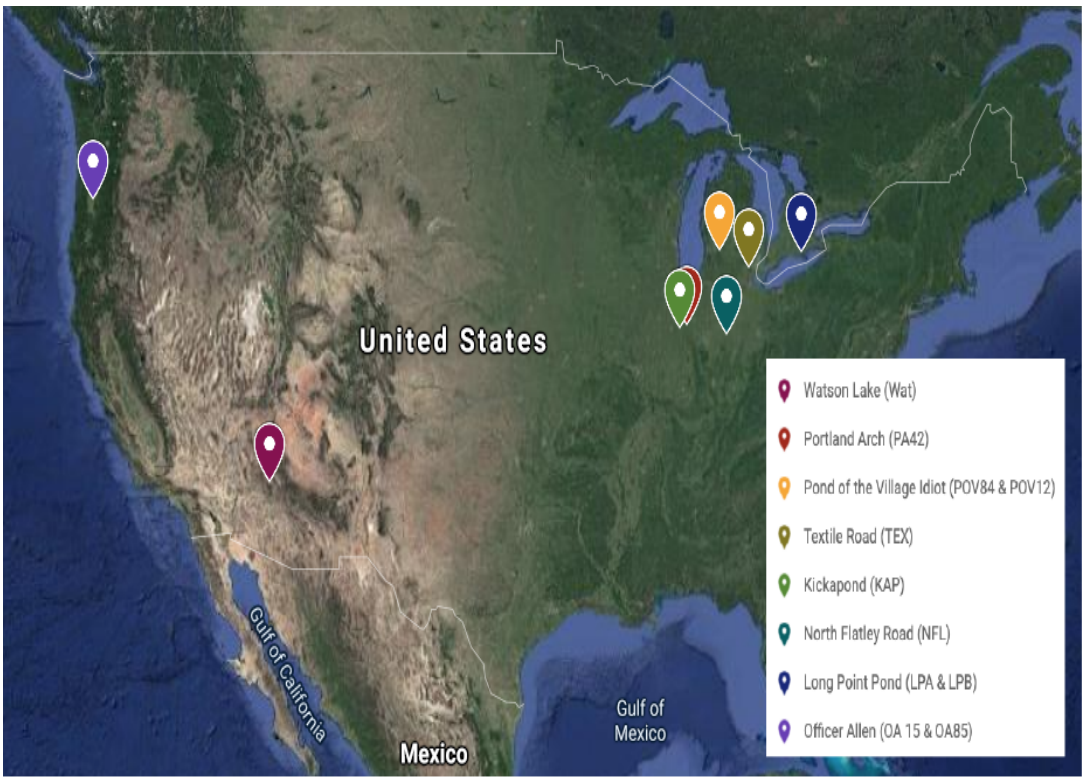

Figure 3. Geographical Distribution of *Daphnia pulex* populations. Map of the locations of the various populations selected during this study. The majority of the populations were extracted from locations in the Midwestern US and Canada, with others congregating in Oregon and Arizona.

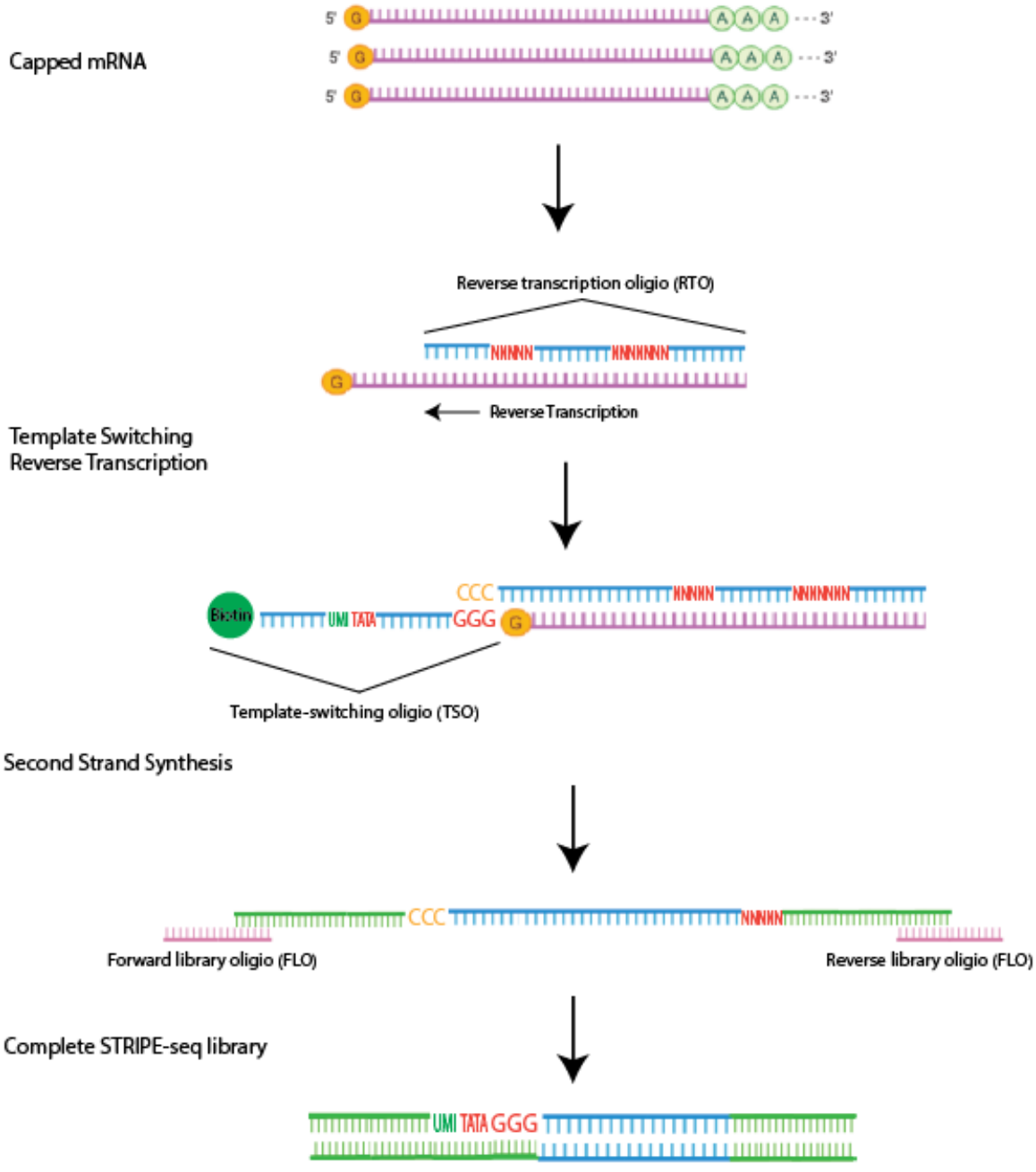

Figure 4. STRIPE-seq protocol. STRIPE-seq uses custom oligonucleotides to perform template switching reverse transcription, using reverse transcriptase (Superscript II Reverse Transcriptase) to construct a cDNA library. Reverse transcriptase naturally synthesizes 3 untemplated C's, which allows for the binding of

the TSO, which is followed by a size-based bead collection to remove and primer dimers. During second strand synthesis, the FLO and RLO bind to the cDNA molecule and are amplified using PCR duplication. Finally, a secondary, more extensive bead cut is conducted to once again remove and primer dimers, which results in a completed STRIPE-seq library ready for Illumina sequencing.

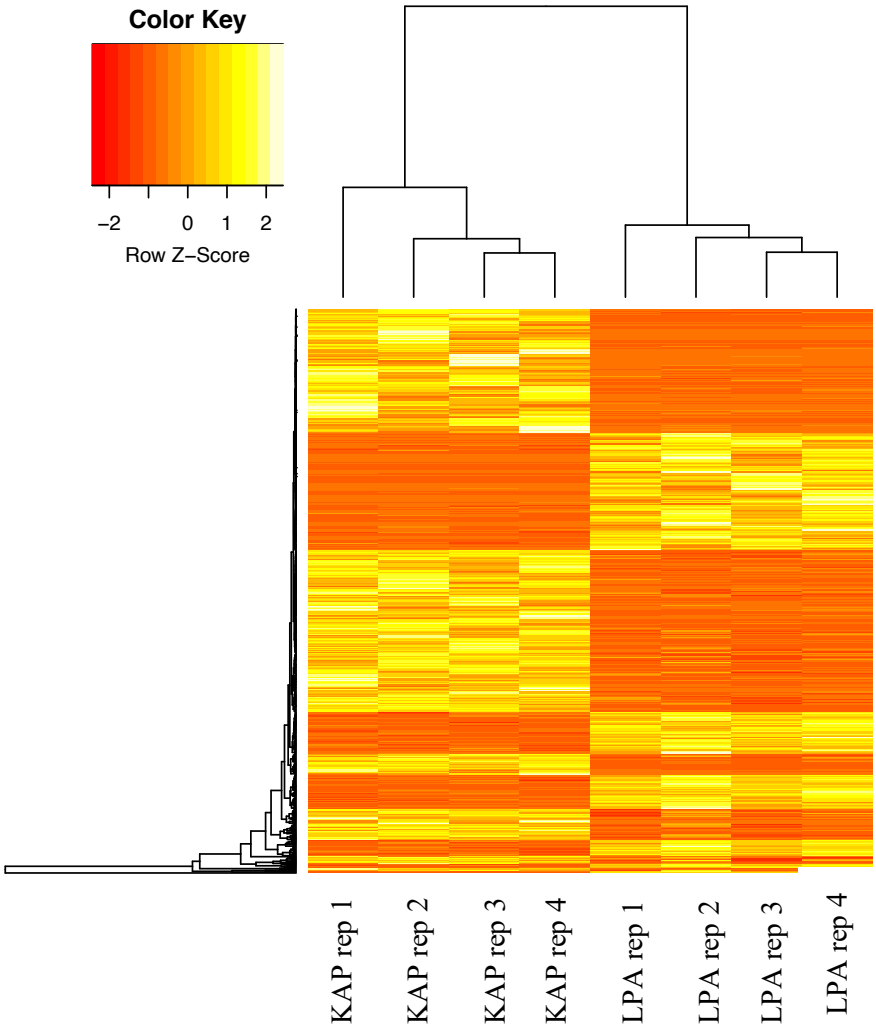

Figure 5. RNA-seq reveals population-level expression differences between D. pulex populations. Results of the RNA-seq analysis reveal the occurrence of differentially expressed genes between arbitrarily selected KAP and LPA populations. Heatmap contains only the known differentially expressed genes within *Daphnia* and indicates conservation between replicates, as well as the presence of gene expression differences across the selected populations.

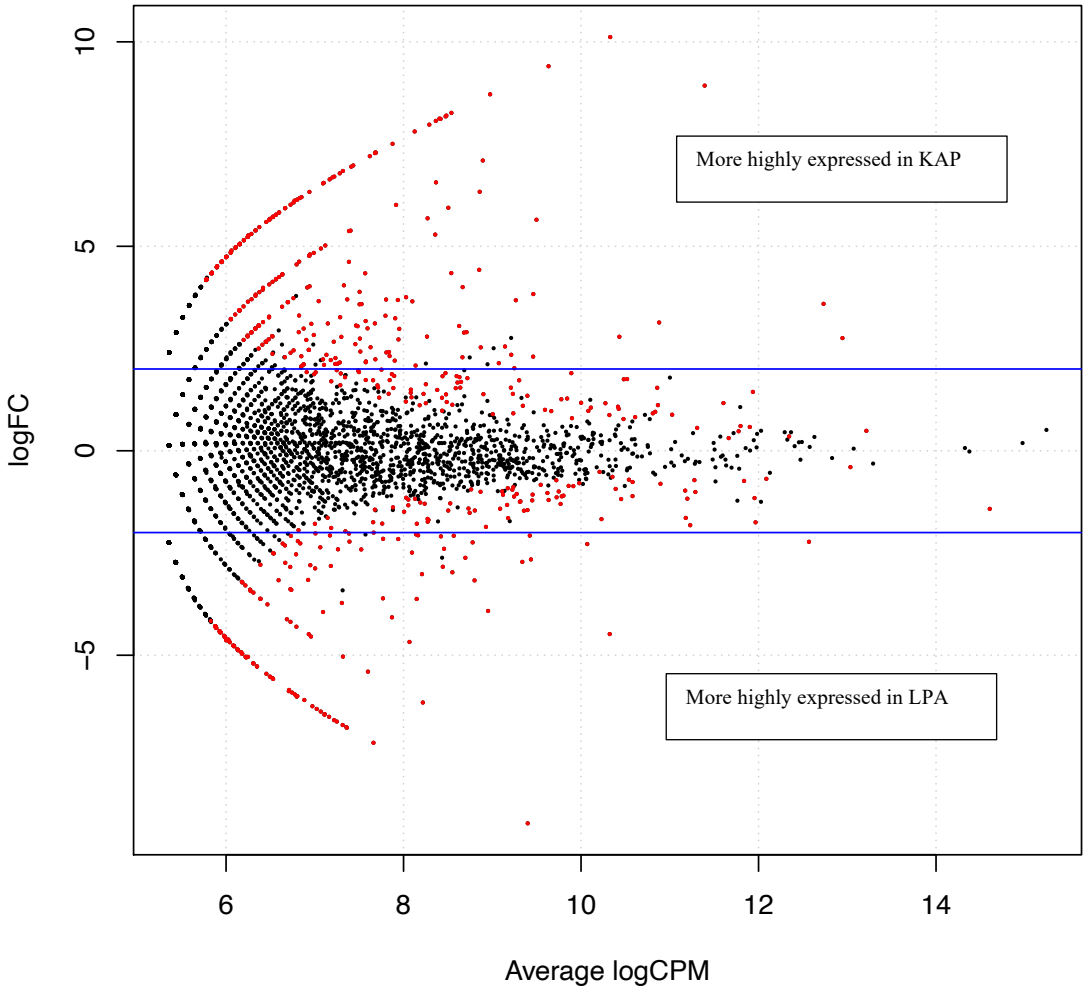

Figure 6. Differentially expressed genes between *D. pulex* populations, LPA and KPA, revealed by RNA-seq. Red dots represent statistically significant genes (p=.05)which are more highly expressed in KAP (top) or LPA (bottom).

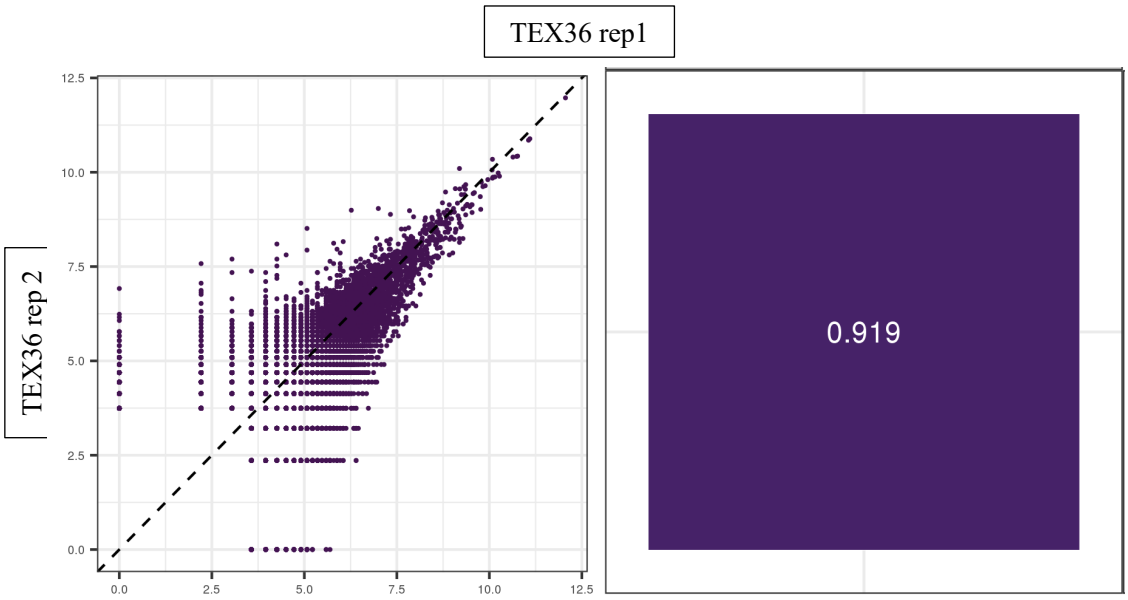

Figure 7. TSS correlation plot of TEX36 replicates. Replicate similarity is validated by the high r^2 value between TEX36 replicate 1 and 2, as illustrated in the top left box. Replicate correlation is an important indicator of standardized replicates, as facilitated by the clonal reproduction of *Daphnia*.

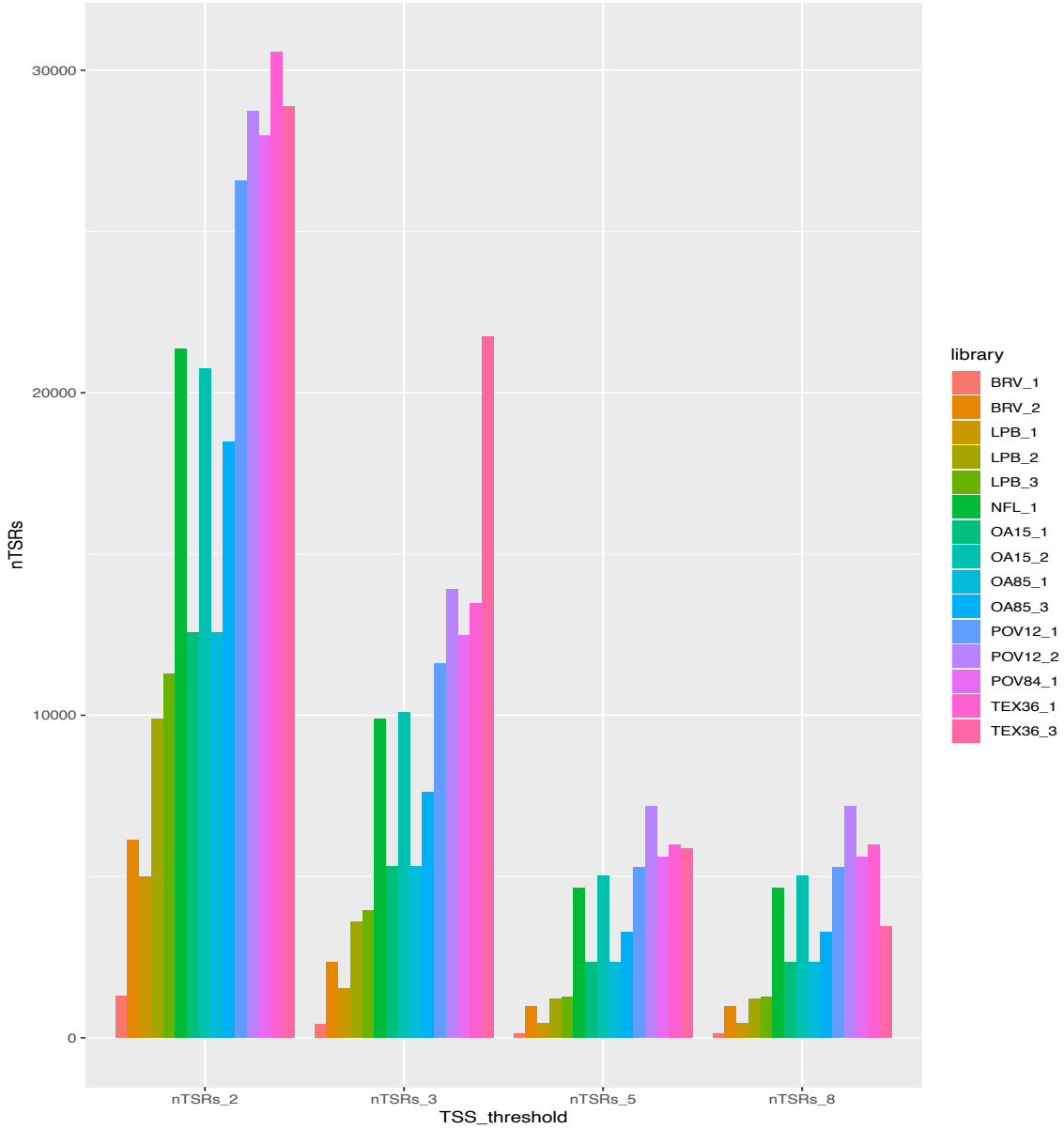

Figure 8. Number of Identified TSRs at a given TSS threshold in each replicate. The plot illustrates the number of identified TSSs that cluster into TSRs at each examined TSS threshold, ranging from 2 to 8. The majority of TSSs cluster into TSRs with a lower TSS threshold, and there is a steady decline as the threshold requirement is raised. The difference in the number of TSRs present at the 5 and 8 TSS threshold is minimized,

suggesting the introduction of an asymptote at a higher threshold. The aligns with the present thinking that most transcription start sites are defined by a lower number of tags, usually between 2-5.

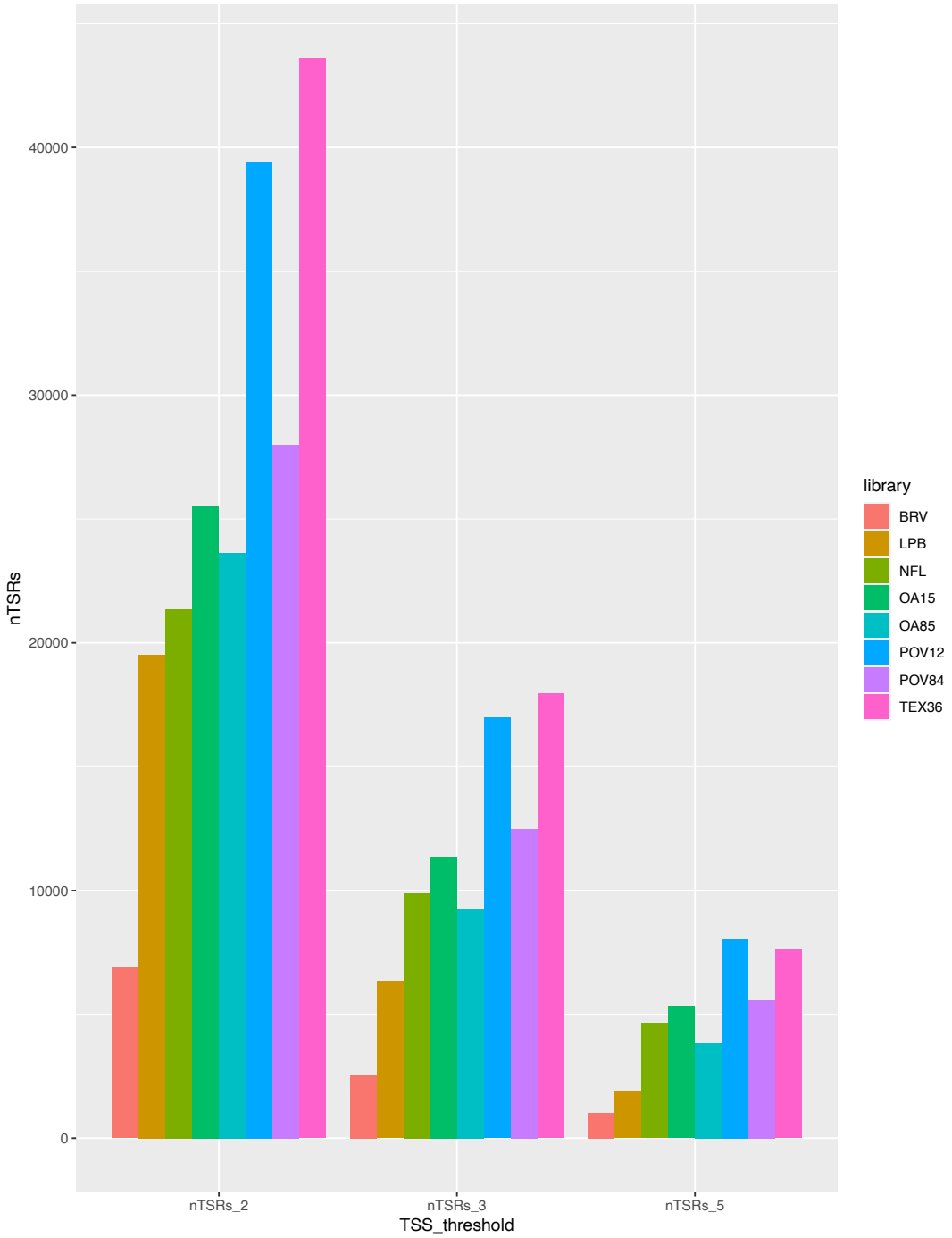

Figure 9. Number of identified TSRs within merged population samples at differing TSS thresholds. The same downward trend is observed as with the replicate data, suggesting a decrease in the identified TSRs as the threshold is increased.

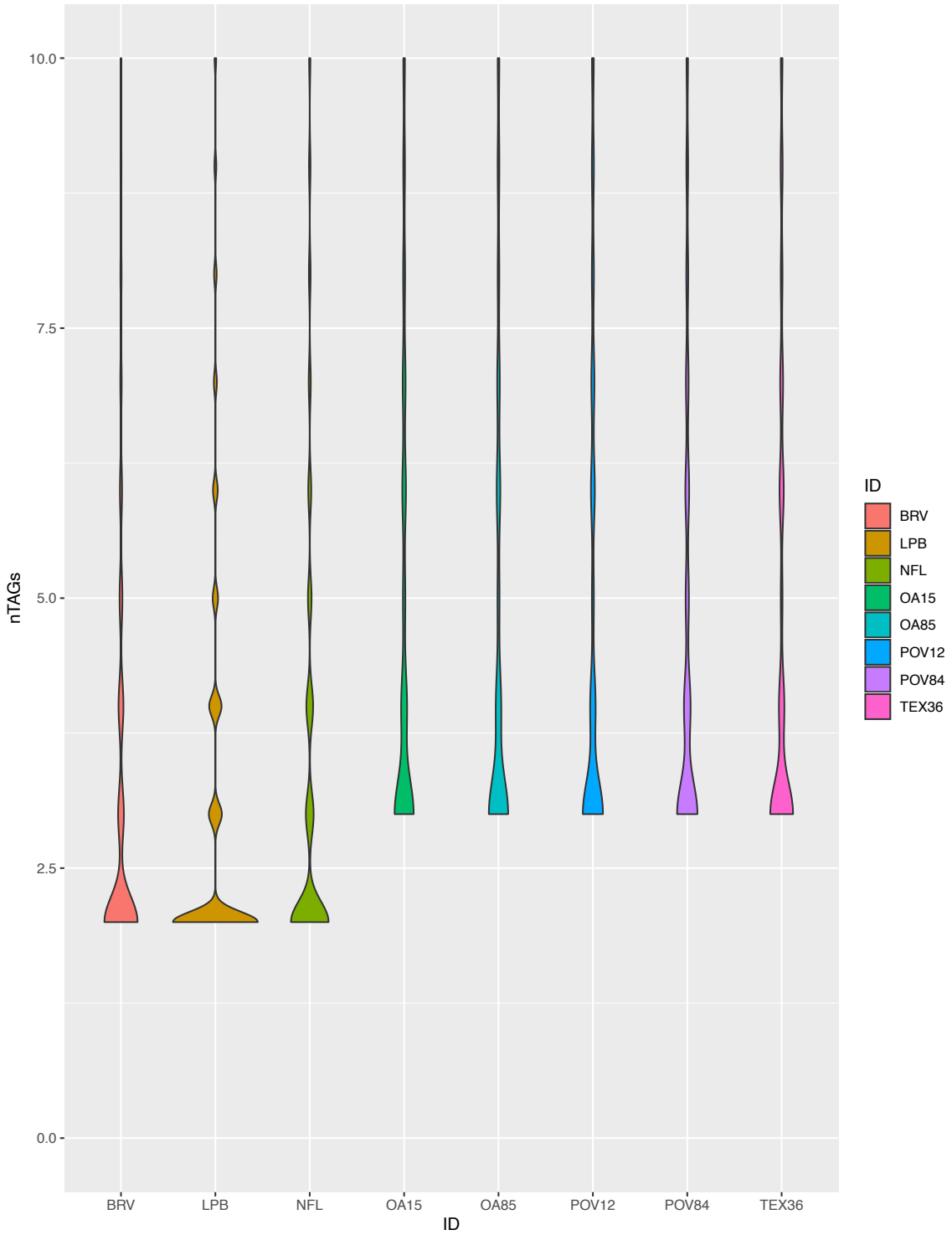

Figure 10. Plot displaying the distribution of the number of tags (nTAGs) present within each merged replicate sample. BRV contains the lowest number of tags, as well as reads (196,258), emphasizing the necessity for read normalization during TSS determination.

Additionally, TEX36, which contains the highest number of tags, also contains the highest number of reads (2,873,988).

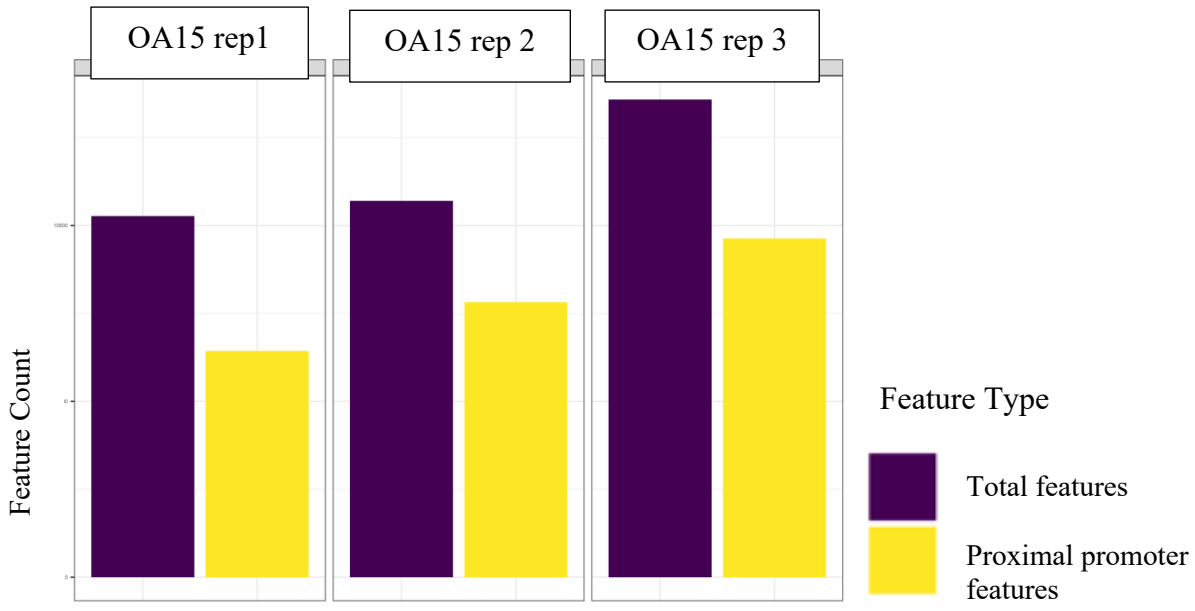

*Figure 11.*Genomic location identification within the proximal promoter of TSSs across three replicates. The majority of TSSs occur within an annotated proximal promoter, which speaks to the quality of the current assembly. In instances when a TSS is not localized within a known promoter region, coupling of the annotation and TSS profiling data can be utilized to impove the quality of the annotation.

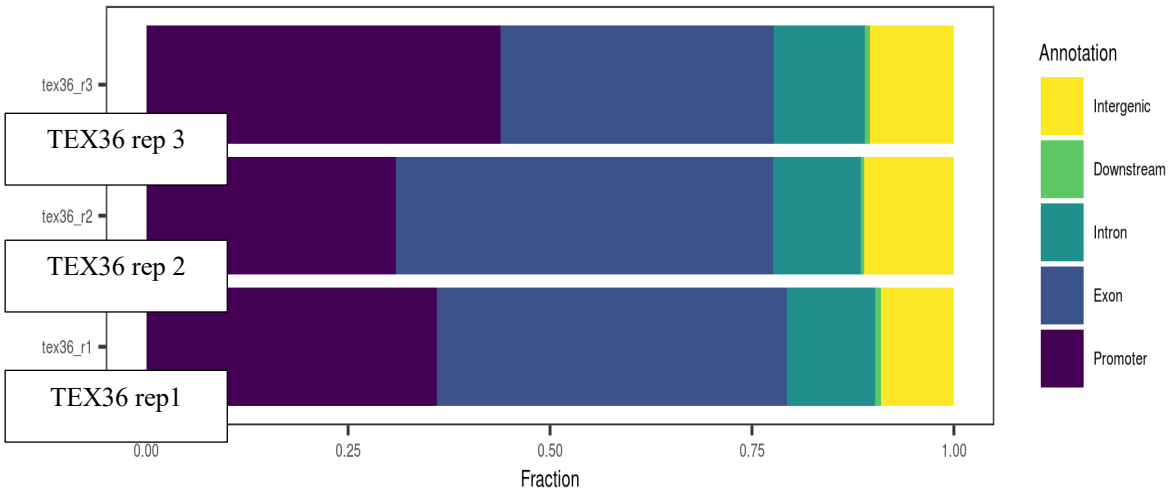

Figure 12. Categorization and relative proportion of the genomic loci of identified TSSs. The majority of identified TSSs are contained within the promoter or the intron, according the most recent annotation; however, there is a subset of TSSs identified within the exon, or intergenic regions. This could be attributed to instances of alternative promoter use, or bi-directional regulatory regions.

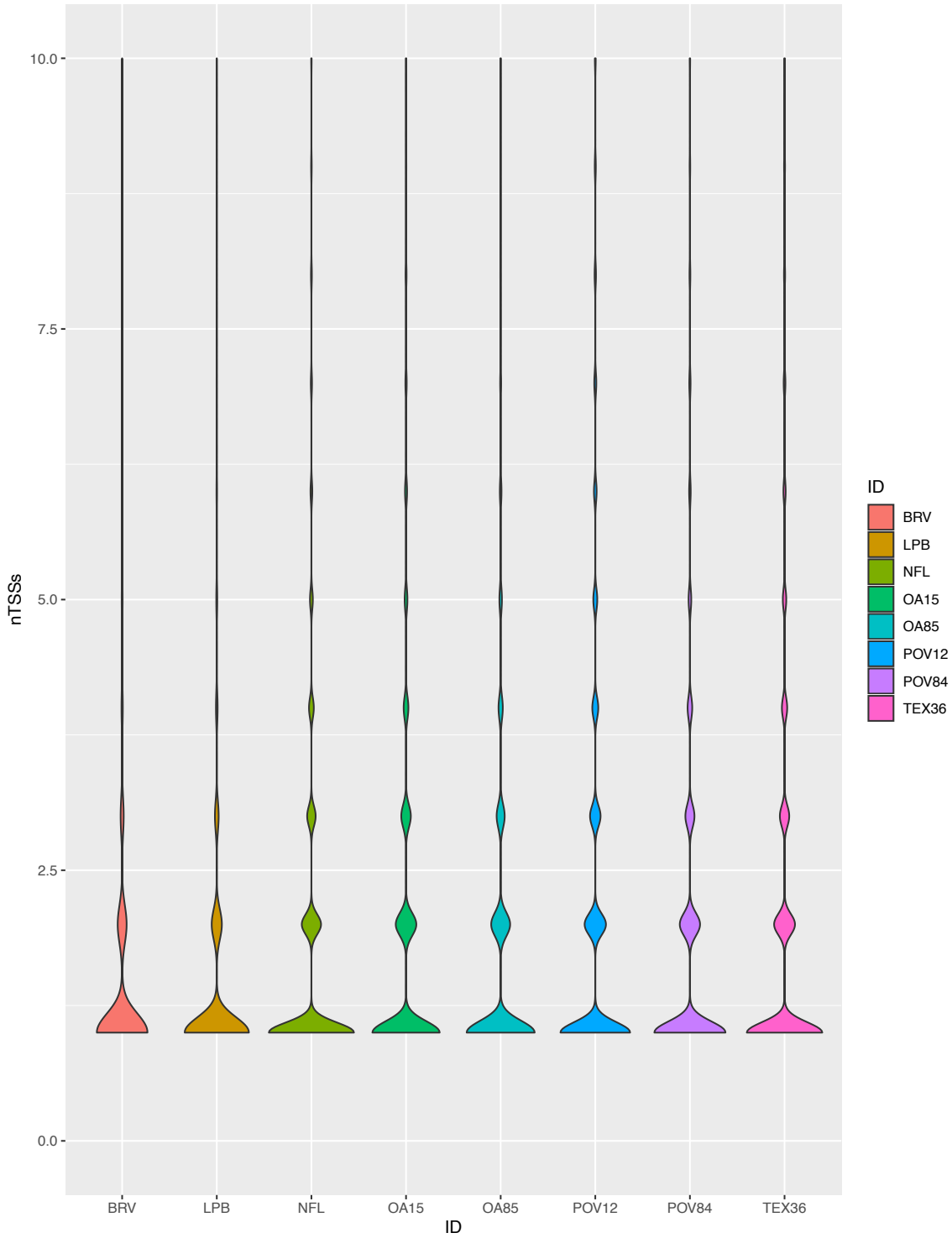

Figure 13. Distribution of identified TSSs depicting a preference towards a lower TSS count. The TSS threshold is the number of tags necessary at a given site to be considered a TSS; transcription initiation within the populations is occuring most often in TSRs

defined by a smaller number of TSSs, suggesting an inflation in transcript abundance associated with tissue-specific genes.

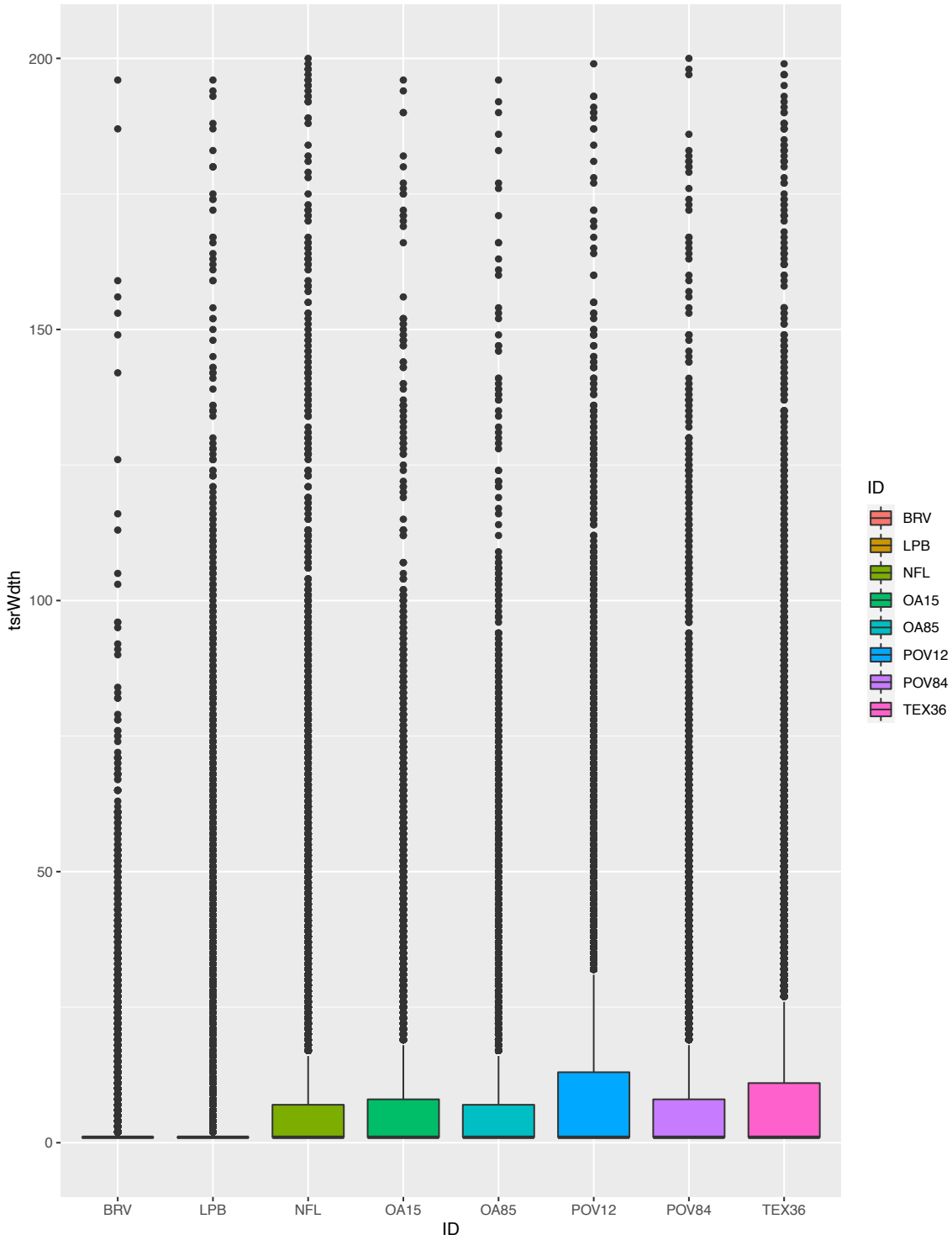

Figure 14. Distribution of the TSR width of selected *D. pulex* populations. TSR width is defined as the absolute value between the coordinate of the first TSS and the last TSS

clustered into a TSR. The majority of TSRs can be linked to a TSR width hovering around lower numbers, which suggests the presence of an abundance of sharp promoters.

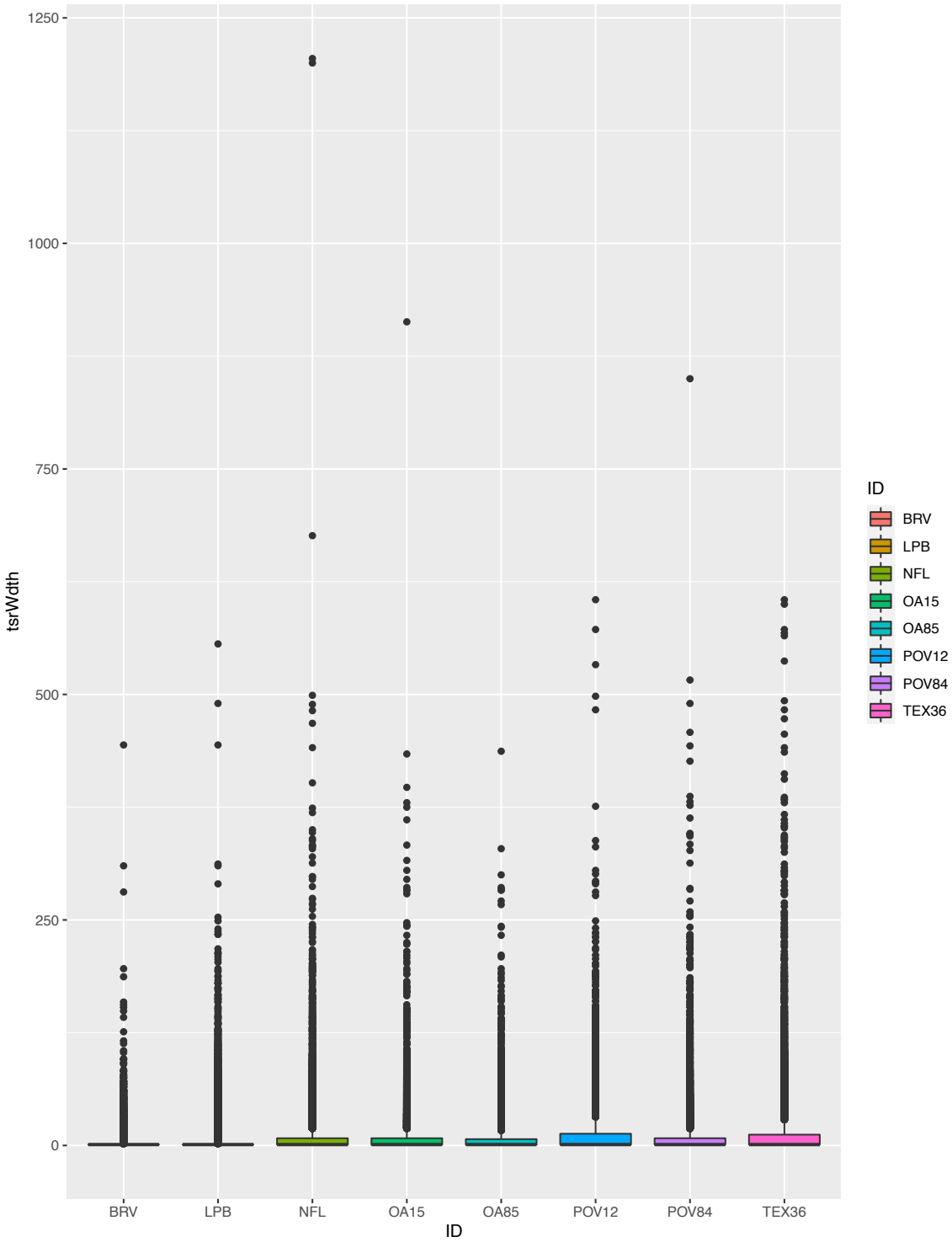

Figure 15. TSR width across various populations suggests the majority of TSRs are less than 250 bp in length; as previously mentioned, the overarching majority of TSRs are

constrained in their genomic distribution, suggesting the presence of peaked, or sharp promoters.

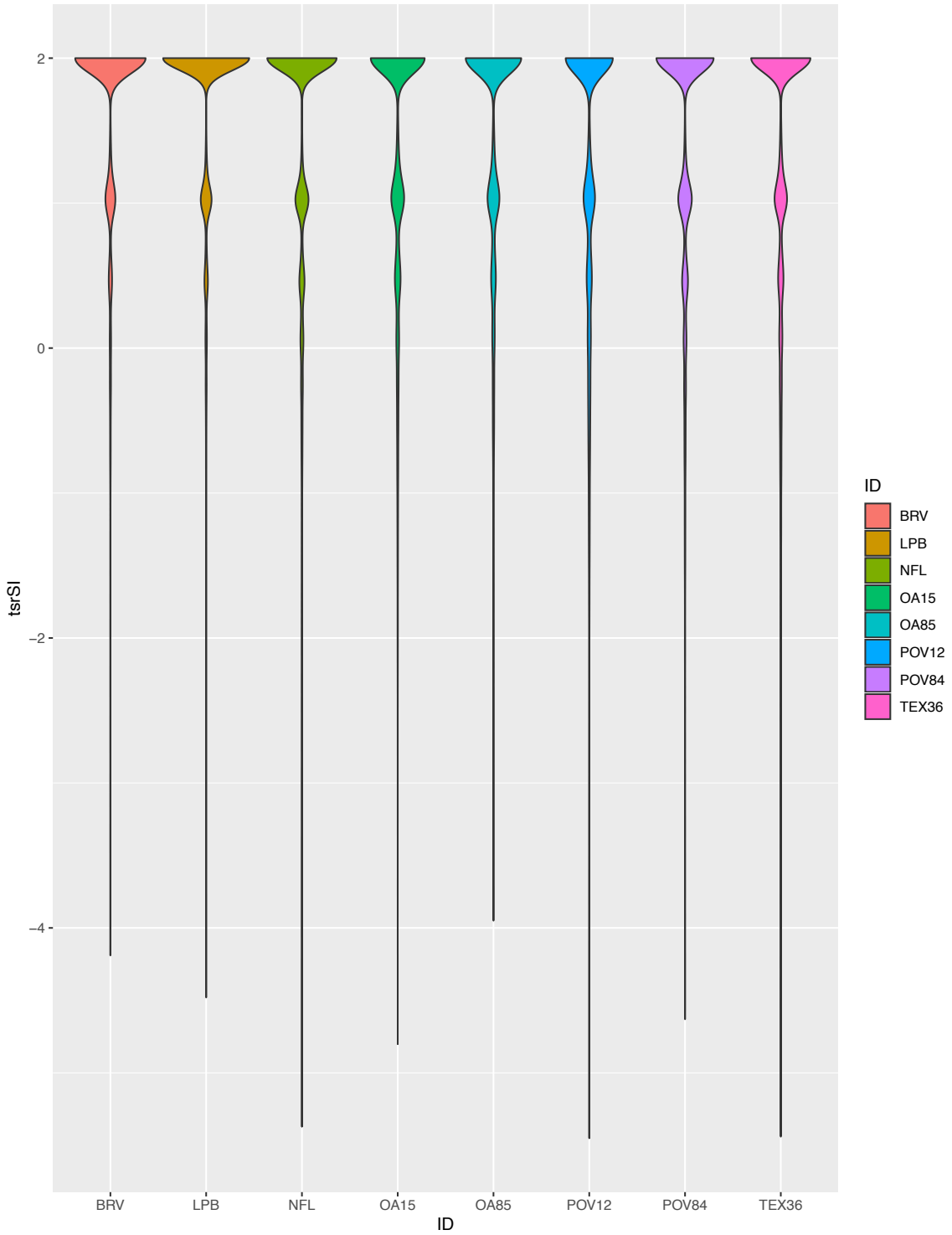

Figure 16. Distribution of promoter architecture as described by the Shape Index. TSR SI is a metric which is correlated with promoter width, and further suggests the presence of bimodal promoter architecture, which is apparent through the separate distribution in the

plot. A SI with a value greater than -1 is classically defined as a peaked promoter, whereas a value less than or equal to -1 is characterized as a broad promoter.

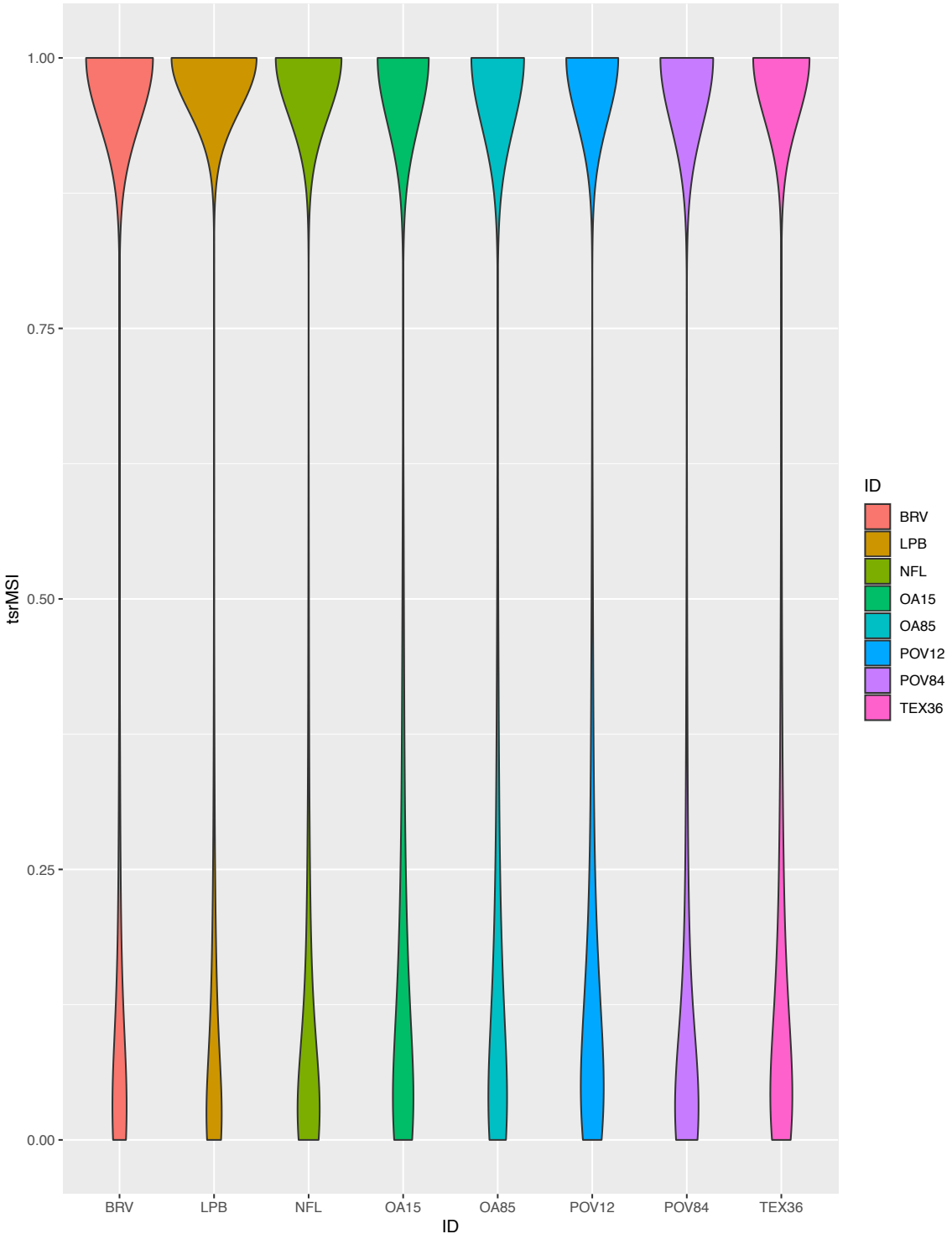

Figure 17. Modified shape index presents bimodal distribution of promoter shape classes.

Because of the assigned boundaries within the mSI metric, mSI values that cluster around

1 represent peaked TSRs and values approaching 0 are characterized by broad promoters, thus promoting the presence of two different promoter classes.

Table 1

Geographical locations of various Daphnia populations

Population	Location	State
BRV	Brookville Reservoir	Indiana
LPB	Long Point Pond	Ontario, Canada
OA15	Office Allen	Oregon
OA85	Officer Allen	Oregon
POV12	Pond of the Village Idiot	Michigan
POV84	Pond of the Village Idiot	Michigan
TEX36	Textile Road	Michigan
NFL36	North Flatley Road	Indiana

Table 2

Quantification of the number of reads and TSS during various point of data analysis, including the raw read count (sequenced libraries), the number of aligned reads (following quality control), and the number of transcription start sites identified. Averages for each category are also included. Overall, an average of 21156324 reads were sequenced, an average of 648883 of these reads were aligned to the genome, resulting in an average of 206396 identified TSSs across populations.

Summary of Identified Read Counts During Various Stages of Data Analysis

Library	Raw Read Count		Number of aligned reads		Number of identified TSSs	
		Average raw read count		Average # of aligned reads		Average TSSs
BRV		12895156	196258	98129		34330
BRV_1	5760382		36466		14383	
BRV_2	20029930		159792		54277	
LPB		24180613.7	324764	193858		67969
LPB_1	21313740		104144		38053	
LPB_2	23691542		220620		75210	
LPB_3	27536559		256810		90643	
OA15		20889423	1318054	658833		199959
OA15_1	20851235		414978		133376	
OA15_2	20927611		902688		266542	
OA85		20577532	1117776	558888		188592
OA85_1	21547485		414978		133376	
OA85_3	19607578		702798		243808	
POV12		29276002	2416626	1207813		376973
POV12_1	21079011		1087658		363900	
POV12_2	37472993		1327968		390046	
TEX36		19451548	2873988	957996		302246
TEX36_1	19171657		1432688		441738	
TEX36_2	19103677		192888		68555	
TEX36_3	20079310		1248412		396444	
NFL_1	20851235		741608		217090	
POV84_1	19477234		1137632		374902	
		21156324		648883		206396

Notes: Samples lacking a merged label represent a single replicate

Table 3. *Number of TSRs present for each TSS threshold.*

library	nTSRs_2	nTSRs_3	nTSRs_5	nTSRs_8
BRV_1	1299	429	134	134
BRV_2	6136	2349	985	985
LPB_1	4996	1541	439	439
LPB_2	9890	3603	1200	1200
LPB_3	11290	3944	1262	1262
NFL_1	21368	9900	4641	4641
OA15_1	12559	5325	2353	2353
OA15_2	20758	10103	5011	5011
OA85_1	12559	5325	2353	2353
OA85_3	18470	7616	3276	3276
POV12_1	26578	11618	5273	5273
POV12_2	28742	13912	7186	7186
POV84_1	27985	12478	5605	5606
TEX36_1	30572	13470	5971	5971
TEX36_3	28869	21744	5873	3444

library	nTSRs_2	nTSRs_3	nTSRs_5
BRV	6874	2542	1028
LPB	19522	6343	1897
NFL	21368	9900	4641
OA15	25497	11364	5328
OA85	23642	9237	3826
POV12	39432	16977	8056
POV84	27985	12478	5605
TEX36	43591	17947	7600

The TSS threshold is defined here as the number of tags required to indicate a TSS; the resulting number of TSRs represents the clustered TSSs that met the described threshold. Both the merged TSS files, and the individual replicates were analyzed at differing thresholds. There were no additional TSRs identified when the threshold was increased from 5 to 8 within the replicates, barring *TEX36_replicate3*. All of the merged samples experienced consistent drops in the number of identified TSRs as the TSS threshold was increased, although the populations were not analyzed at a threshold of 8 due to the lack of differences observed within the replicates.

Table 4

Normalization of the number of reads (in millions) and the determination of the appropriate TSS threshold

Library	8/nreads (millions)	5/nreads (millions)	3/nreads (million)	2/nreads (million)	TSS Threshold
BRV_merged	N/A	25.47666847	15.28600108	**10.19066739**	2
BRV_1	219.3824384	137.114024	82.26841441	54.84560961	
BRV_2	50.06508461	31.29067788	18.77440673	12.51627115	
LPB	N/A	8.597358204	**5.158414922**	3.438943282	3
LPB_1	76.81671532	48.01044707	28.80626824	19.20417883	
LPB_2	36.26144502	22.66340314	13.59804188	9.065361255	
LPB_3	31.151434991	19.46964682	11.68178809	7.787858728	
OA15_merged	N/A	**3.793471284**	2.276082771	1.517388514	5
OA15_1	19.278130441	12.0488315	7.229298903	**4.819532602**	
OA15_2	8.862419795	**5.539012372**	3.323407423	2.215604949	
OA85_merged	N/A	**4.473168148**	2.683900889	1.789267259	3
OA85_1	19.278130441	12.0488315	7.229298903	**4.819532602**	
OA85_3	11.383071666	7.114419791	**4.268651874**	2.845767916	
POV12_merged	N/A	2.06985684	**1.241914104**	0.827942736	5
POV12_1	7.355253214	**4.597033259**	2.758219955	1.838813303	

POV12_2	6.024241548	**3.765150967**	2.25909058	1.506060387	
TEX36_merged	N/A	1.739742824	**1.043845695**	0.69589713	3
TEX36_1	**5.583909407**	3.489943379	2.093966027	1.395977352	
TEX36_2	41.47484551	25.92177844	15.55306706	10.36871138	
TEX36_3	6.408140902	**4.005088064**	2.403052838	1.602035226	
NFL_1	10.787370155	6.742106342	**4.045263805**	**2.696842537**	3
POV84_1	7.032150994	**4.395094371**	2.637056623	1.758037749	5

Notes: Samples lacking a merged label represent a single replicate

Table 5

Summary of number of tags, TSSs, and TSRS, including relevant averages.

Library	Total number of tags	Average number of tags per TSS	Number of identified TSSs	Number of identified TSRs	Average number of TSSs/TSR
BRV	39795	1.16	34330	6874	4.99
LPB	198468	2.92	67969	19522	3.48
OA15	251267	1.25	199959	11364	17.60
OA85	170037	1.11	188592	9237	20.42
POV12	441537	1.17	376973	16977	22.20
TEX36	510045	1.69	302246	17947	17.12
NFL	198468	0.94	217090	21368	10.16
POV84	185341	0.49	374902	12478	30.05
Average	249369	1.34	220257.625	14471	15.75

Note: NFL and POV84 are single replicates

REFERENCES

Balwierz, P. J., Carninci, P., Daub, C. O., Kawai, J., Hayashizaki, Y., Van Belle, W., Beisel, C., & van Nimwegen, E. (2009). Methods for analyzing deep sequencing expression data: Constructing the human and mouse promoterome with deepCAGE data. *Genome Biology*, *10*(7). https://doi.org/10.1186/gb-2009-10-7-r79

Batut, P., & Gingeras, T. R. (2013a). RAMPAGE: Promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. *Current Protocols in Molecular Biology*, *104*(SUPPL.104). https://doi.org/10.1002/0471142727.mb25b11s104

Batut, P., & Gingeras, T. R. (2013b). RAMPAGE: Promoter activity profiling by paired-end sequencing of 5'-complete cDNAs. *Current Protocols in Molecular Biology*, *104*(SUPPL.104). https://doi.org/10.1002/0471142727.mb25b11s104

FANTOM Consortium, F., Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., Hoon, M. J. L. de, Haberle, V., Lassmann, T., Kulakovskiy, I. V, Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. https://doi.org/10.1038/nature13182

Corp., Z. R. (n.d.). *Direct-zol™ RNA Microprep*. Retrieved July 6, 2020, from https://files.zymoresearch.com/protocols/_r2060_r2061_r2062_r2063_direct-zol_rna_microprep.pdf

Gregory, B. D. (2018). Shedding some blue light on alternative promoter usage in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(30), 7654–7656. https://doi.org/10.1073/pnas.1809312115

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, *19*(10), 621–637. https://doi.org/10.1038/s41580-018-0028-8

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., Booth, B. W., Zhang, D., Wan, K. H., Yang, L., Boley, N., Andrews, J., Kaufman, T. C., Graveley, B. R., Bickel, P. J., Carninci, P., Carlson, J. W., & Celniker, S. E. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research*, *21*(2), 182–192. https://doi.org/10.1101/gr.112466.110

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. M., & Kadonaga, J. (2008). The RNA Polymerase II Core Promoter – the Gateway to Transcription. *Current Opinion Cell Biology*. https://doi.org/10.1038/jid.2014.371

Juven-Gershon, T., Hsu, J. Y., & Kadonaga, J. T. (2008). Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes and Development*, *22*(20), 2823–2830. https://doi.org/10.1101/gad.1698108

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2009). TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics*.

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, *13*(4), 233–245. https://doi.org/10.1038/nrg3163

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Lynch, M. (1983). ECOLOGICAL GENETICS OF *DAPHNIA PULEX*. *Evolution*, *37*(2), 358–374. https://doi.org/10.1111/j.1558-5646.1983.tb05545.x

Nelson, J. D., Denisenko, O., & Bomsztyk, K. (2006). Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nature Protocols*, *1*(1), 179–185. https://doi.org/10.1038/nprot.2006.27

Pipkin, M., & Lichtenheld, M. (2006). *A reliable method to display authentic DNase I hypersensitive sites at long-ranges in single-copy genes from large genomes*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1388096/

Policastro, R. A., Raborn, R. T., Volker, B. P., & Zentner, G. E. G. (2020). *Simple and efficient measurement of transcription initiation and transcript levels with STRIPE-seq*. *2011*(2865), 1–9.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., & Sandelin, A. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biology*, *7*(8). https://doi.org/10.1186/gb-2006-7-8-r78

Raborn, R. T., Sridharan, K., & Volker, P. B. (2017). TSRchitect: Promoter identification from large-scale TSS profiling data. *Bioconductor*.

Rach, E. A., Yuan, H. Y., Majoros, W. H., Tomancak, P., & Ohler, U. (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology*, *10*(7). https://doi.org/10.1186/gb-2009-10-7-r73

Rapid amplification of 5′ complementary DNA ends (5′ RACE). (2005). *Nature Methods*, *2*(8), 629–630. https://doi.org/10.1038/nmeth0805-629

Reyes, A., & Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, *46*(2), 582–592. https://doi.org/10.1093/nar/gkx1165

Robinson, M., McCarthy, D., & Smyth, G. (2010). *Empirical Analysis of Digital Gene Expression Data in R*.

Saito, T. L., Hashimoto, S. I., Gu, S. G., Morton, J. J., Stadler, M., Blumenthal, T., Fire, A., & Morishita, S. (2013). The transcription start site landscape of C. elegans. *Genome Research*, *23*(8), 1348–1361. https://doi.org/10.1101/gr.151571.112

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. *Nature Reviews Genetics*, *8*(6), 424–436. https://doi.org/10.1038/nrg2026

Schmidt, W. M., & Mueller, M. W. (1999). CapSelect: A highly sensitive method for 5′ ′ ′ ′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. In *Nucleic Acids Research* (Vol. 27, Issue 21).

Schor, I. E., Degner, J. F., Harnett, D., Cannavò, E., Casale, F. P., Shim, H., Garfield, D. A., Birney, E., Stephens, M., Stegle, O., & Furlong, E. E. M. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*, *49*(4), 550–558. https://doi.org/10.1038/ng.3791

Shannon, C. E. (1948). A Mathematical Theory of Communication. In *The Bell System Technical Journal* (Vol. 27).

Stollewerk, A. (2010). The water flea Daphnia - A “new” model system for ecology and evolution? In *Journal of Biology* (Vol. 9, Issue 2, p. 21). BioMed Central. https://doi.org/10.1186/jbiol212

Tang, D. T. P., Plessy, C., Salimullah, M., Suzuki, A. M., Calligaris, R., Gustincich, S., & Carninci, P. (2012). *Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching*. https://doi.org/10.1093/nar/gks1128

Ye, Z., Molinier, C., Zhao, C., Haag, C. R., & Lynch, M. (2019). Genetic control of male production in Daphnia pulex. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(31), 15602–15609. https://doi.org/10.1073/pnas.1903553116

Zhang, S. X. L., Searcy, T. R., Wu, Y., Gozal, D., & Wang, Y. (2007). Alternative promoter usage and alternative splicing contribute to mRNA heterogeneity of mouse monocarboxylate transporter 2. *Physiological Genomics*, *32*(1), 95–104. https://doi.org/10.1152/physiolgenomics.00192.2007