Title
Methods for the Analysis of High Throughput Sequencing Data

Author
Boley, Nathan

Publication Date
2014

Peer reviewed|Thesis/dissertation

Methods for the Analysis of High Throughput Sequencing Data

By

Nathan Paul Boley

A thesis submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Commitee in Charge:

Professor Peter J. Bickel, Chair

Professor Haiyan Huang

Professor Ian Holmes

Doctor Susan Celniker

Fall 2014

Abstract

Methods for the Analysis of High Throughput Sequencing Data

by

Nathan Paul Boley

Doctor of Philosophy in BioStatistics

University of California, Berkeley

Professor Peter J. Bickel, Chair

As the cost of short read DNA sequencing continues to drop, new experiments are being developed which allow scientists to probe subtle biological phenomenon on a genome wide scale. These new experiments bring new analytical challenges in the form of large amounts of noisy data. Here we present models and tools for the analysis of high throughput sequencing data.

Contents

CHAPTER 1

Introduction

The development of technologies which are able to cost effectively sequence hundreds of millions of short DNA fragments has led to a revolution in experimental genetics. High throughput sequencing technologies provide a common platform through which researchers can interrogate diverse biological features including DNA sequence, mRNA sequence, RNA structure, epigentic modifications, and chromatin structure. Obtaining useful information from the resulting mounds of data can be challenging, particularly since technology development has been driven by what is technically feasible, often with little regard for experimental design. To infer relationships between sequenced reads and biological events, statistical methods must account for the large amounts of noise from technical and biological sources, must contend with possible identifiability problems, and should be computationally feasible for large datasets.

Here we present several methods and associated tools which address the challenges involved with the analysis of data from high throughput sequencing experiments. In the first chapter, we present a careful treatment of short read mapping and the process of associating RNA fragments with the locus from which they originated. We show that by applying these models we are better able to distinguish between competing biological hypotheses than current methods. In the second, we show that by integrating data from multiple experiment types, we are better able to identify noise. In the third, we present a model for the quantification of RNA elements, and derive conditions under which the quantity of such elements is estimable from short read sequencing technologies. Finally, in chapter 4, we present an integrated model for the discovery of RNA elements, and apply it to data collected by the modENCODE consortium.

CHAPTER 2

Statmap

2.1. Introduction

For organisms with a reference genome [16], analysis of sequencing data can be substantially simplified and improved by comparing the sequenced reads to the reference genome. This chapter focuses on this process of using a reference genome to inform the analysis of sequencing data.

Analysis of sequencing experiments performed on an organism with a reference genome typically begins by associating each read with one or more locations in the genome from which the read could have originated, a process known as "mapping" or "read alignment". For example, if one is attempting to identify novel single nucleotide mutatations in a population, a typical analysis pipeline would involve aligning the sequenced reads from each individual to the reference genome, and then identifying single bases that differ between the reference genome and the sequenced reads. Conceptually the goal is simple: to identify bases that differ from the reference sequence. However, this analysis is confounded by noise introduced during sample preparation, noise in the sequencing technologies, and an identifiability problem stemming from duplicated sequences in the reference genome.

If there was no mapping ambiguity or sequencing error, one would be able to identify the location where a read originated by finding the genomic location with a sequence that matches the read's sequence. In reality, the sequence reported by the sequencing technology might not match any genomic location. Base level mutations, insertions, and deletions can be introduced during the steps leading up to the DNA's purification, during the PCR amplification steps that typically preceed sequencing, or during the sequencing process. In a typical RNA experiment performed using the newest Illumina sequencing and sample preparatory technologies, bases on average have roughly a 5% chance of being reported incorrectly. Thus, in a read that is 150 basepairs long, we expect to have 7-8 mutations on average; less than 0.1% of reads will perfectly match the sequence from which they originated.

A single read may also map to several genomic locations. If the distribution of bases within a genome were uniform, then reads of 150 bases would be long enough to uniquely map every read with high probability. However, new genomes originate not only by single base level

changes but by large duplications and re-arrangments. Under the uniform base composition model, the probability that a 150 basepair segment taken randomly from the human genome would match another segment with less than 7 mutations is $< 1 \times 10^{-76}$; however, in the GRCh38 reference genome, 18% of 150 basepair fragments match another region with fewer than 7 mutations. Although the cost per base of read sequencing has been rapidly falling, read lengths are increasing slowly and often at the cost of accuracy. Furthermore, because of the structure of genomic sequences, moderate increases in read length do little to improve the identifiability problem.

Because of these problems, drawing strong biological conclusions from a single read is very difficult. Fortunately, modern sequencing experiments produce tens of millions of reads, which can be jointly modeled to improve mapping accuracy. Such modelling is the focus of this paper.

Good models of sequencing experiments are complex, and typically assay specific. Furthermore, good techniques must balance accuracy and computational cost. Thus, for pedagogical reasons, we develop the model in three sections. In the first, we present the full model which, although general enough to apply to any sequencing experiment, has more parameters than can be fit without additional assumptions. Next, we discuss the assay specific parameterizations and the corresponding estimation procedures. Finally, we discuss approximations which are necessary to make the estimation procedure feasible.

2.2. Model

Sequencing involves randomly selecting a pre-determined number DNA fragments from a solution and measuring their sequence. The output of such experiments is a list of genomic sequences. Each sequence in this list is referred to as a read; we use r_i to indicate the observed sequence of the i'th read. We assume that each observed read originated from a single location in the reference genome. We use the term "map" to refer to the process of inferring the genomic location at which a read originated.

2.2.1. Likelihood Model. Since the number of fragments in solution is much larger than the number of fragments sequenced, we model sequencing as a simple random sample with replacement. We define two model parameters.

The first, $P[r_i|g_{j,l}]$, is the conditional probability of observing r_i given that it originated from a fragment of length l that originated from genomic location j. This specifies the sequencing error model. We assume that the probability of observing r_i is only dependent on genomic location through the location's sequence. That is, for two locations j and j' with the same

sequence, we assume that $P[r_i|g_{j,l}] = P[r_i|g_{j',l}]$. We discuss further paramaterizations in 2.3.1.

The second, $P[g_{j,l}]$, is the marginal probability of a sampling a fragment of length l originating from genomic locations j. With assay specific knowledge, we can constrain the form of the$P[g_{j,l}]$'s and thus make use of known structure. For example, ChIP-Seq experiments are designed such that observing a fragment at position j with length l implies that the protein of interest was bound somewhere between bases j and $j + l$. Thus, if it were known that a protein were bound at, say, position $j+10$, we would be equally likely to observe a fragment of length l originating from position j. In general the distribution of $P[g_{j,l}]$ is a function of the underlying assay type, the fragment length distribution, and a hidden biological parameter. For instance, in a ChIP-seq experiment, the hidden biological parameter is the binding sites and their relative occupancies.

To formalize this we constrain $P[g_{j,l}]$ to be equal to $\sum_k \theta_k \psi_{f_l}(j, l|k)$ where

f_l: the marginal probability of sequencing a fragment of length l, where $\sum_l f_l = 1$.
θ_k: the marginal probability of sampling read having arisen from biological event k.
$\psi_{f_l}(j, l|k)$: the probability that event k emits a fragment that spans bases j to $j + l$.

We assume that the form of the function $\psi_{f_l}(j, l|k)$ is known from how the assay was designed; we derive the specific form for various assays in 2.3.3.

Then the joint log likelihood of observing a set reads is

$$lhd[\vec{\theta},\, P[r_i|g_{j,l}],\, \psi_{f_l};\, \vec{r},] = \prod_i \sum_j P[r_i|g_{j,l}]P[g_{j,l}] \tag{2.2.1}$$

$$= \prod_i \sum_{j,l} P[r_i|g_{j,l}] \sum_k \theta_k \psi_{f_l}(j, l|k) \tag{2.2.2}$$

where we assume that each read originated due to some biological event and that each biological event produced a fragment, i.e.$\sum_{j,l} P[r_i|g_{j,l}] = 1\ \forall i, l$, $\sum_k \theta_k = 1$, and $\sum_{j,l} \psi_{f_l}(j, l|k) = 1\ \forall k$.

2.3. Parameter Estimation

The likelihood model relies on the estimation of three distinct quantites:

$P[r_i|g_{j,l}]$: the sequence error model

f_l: the fragment length distribution

θ_k: the fraction of fragments in the population that resulted from biological event k

Estimating θ_k is the primary goal of a sequencing experiment, but it is also the most difficult to estimate because it is high dimensional and dependent upon genomic location. For instance, in a ChIP-Seq experiment, the protein of interest typically binds to roughly 100,000 different genomic locations, which means that it is common to estimate a particular binding site's occupancy from less than 100 reads.

In contrast, $P[r_i|g_{j,l}]$ and f_l have the same structure genome wide, are independent of one another, and realistic parameterizations never have more than 1000 parameters. For these reasons we first estimate these quantities, plug them into (2.2.2), and then maximize the likelihood over θ. We describe each of these procedures below.

2.3.1. Sequencing Error Model Estimation. Read errors are introduced during multiple experimental stages including sample preparation, PCR amplification, and sequencing. Sequencing platforms provide an estimate of the sequencing error rate, but this estimate can only account for noise in the sequencing and is thus an underestimate of the true error rate (see Figure 2.3.1)

In addition to the sequencing platforms estimated error rate, position within the read, base type, read pair, and sequenced direction are all correlated with the true error rate. Furthermore, we expect the marginal error rate across positions and adjacent sequencer estimated quality codes to be smooth.

We model mutation events as independent conditional on read position, sequencer estimated error rate, and base type, conditional on read pair and direction. That is, for each read pair-direction combination and given an observed base $b \in [ACGT]$ in position i with sequencing error estimate e_s, we model the probability of observing the base b given that the true base was o by

$$P[b = o|i, e_s, d] = \beta_{b,o,d,0} + \beta_{b,o,d,1} s(i) + \beta_{b,o,d,2} s(e_s) + \beta_{b,o,d,3} \left(s(i) : s(e_s)\right) \quad (2.3.1)$$

where s is an adaptive spline smoothing function.

To estimate the sequencing error model parameters we require sequenced reads with known true underlying sequences. When the experimental design includes "spike-ins" – known sequence combined with the sample to assist in quality control – we can use reads that map to this sequence to estimate the error rates. When we do not have access to such a training set we use reads that only map "well" under the edit distance metric to "unique" sequence within the reference genome. Specifically, we identify regions that are at least 2 kb long

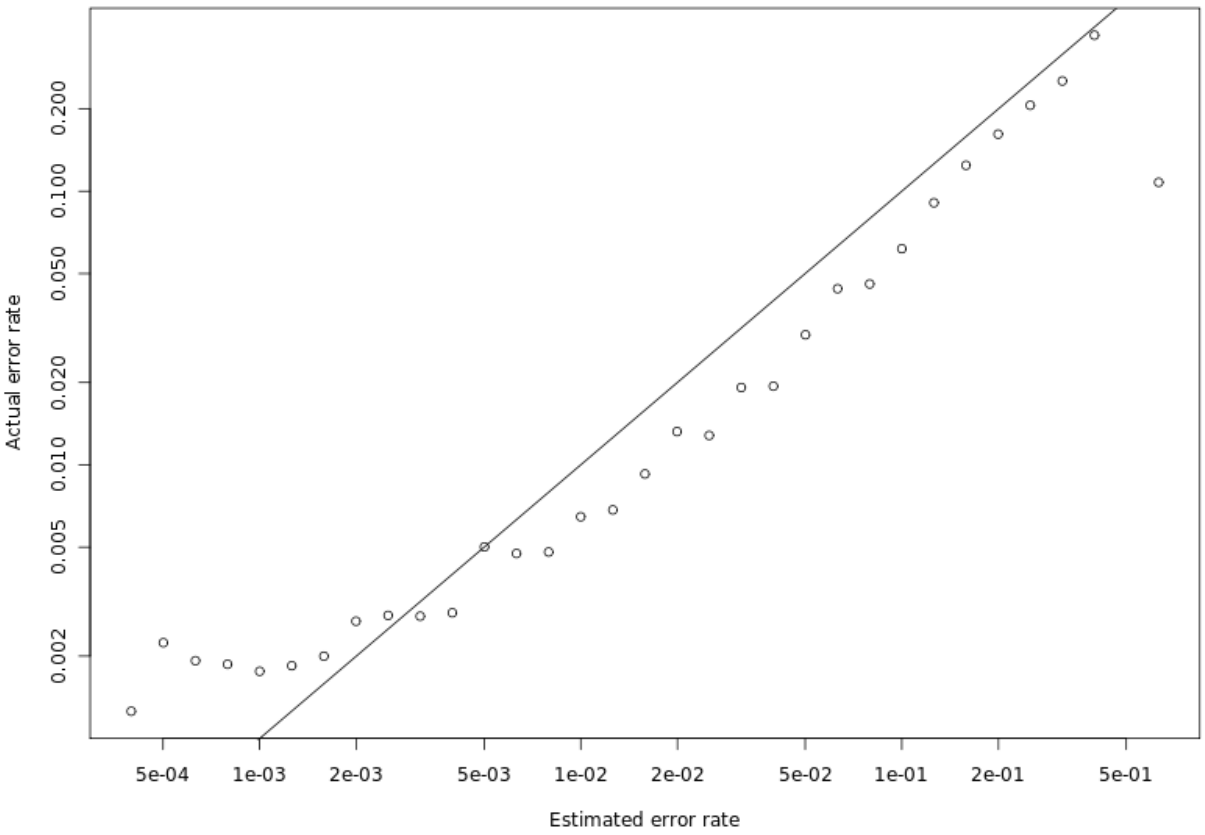

FIGURE 2.3.1. Illumina sequencing technology estimated error rate versus observed for a ChIP-seq experiment involving 8 rounds of PCR. Note that the observed error rate reaches a minimum of roughly 2e-3, which is consistent with the commonly reported PCR error rate of 3e-4 mutations per round.

in which no read length long subsequence maps to another genomic location with less than a 20% mismatch rate. We then map a subset of reads to this subset of the genome using edit distance, assuming that the mapping location with the smallest edit distance is correct. We smooth the marginal distribution of the position- and sequencing technology-dependent reported error rates, and then iterate between fitting a multinomial glm, and re-smoothing. We use the adaptive spline smooth technique suggested in [25] (see 2.3.1).

Some have reported additional errors in RNA-seq assays which are due to random hexamer priming bias Hansen et al. [23]. However, modern RNA-seq protocols use a higher concentration of random primer than the studies in which such biases were identified. We analyzed 112 modENCODE RNAseq data sets (see Chapter 5) and were not able to identify such bias. Because the bias appears to be corrected in modern protocols, we have decided not to model this in statmap. Similarly, [45] reports serial correlation in mutation rates, which

the authors attempt to correct by fitting a first order markov model. When we analyed 112 modENCODE RNAseq data sets, we did not find a significant serial correlation and so do not attempt to model it.

2.3.2. Fragment Length Distribution Estimation. We estimate the fragment length distribution in a similar manner to how we estimate the sequencing error model. We identify genomic regions that are unique, in the sense that most similar alternate genomic sequence is far apart in edit distance. We align reads to these locations, and then use the gaussian kernel smoothed fragment length density with the kernel width set to the standard deviation of the data, as suggested in [25].

2.3.3. Emitted Fragment Density Estimation. Our ability to estimate θ in 2.2.2 depends upon a biological event k's emitted fragment distribution $\psi_{f_l}(l, j|k)$. This function is assay dependent; we have implemented kernels for two common assays which we describe here.

2.3.3.1. *RNA-seq kernel.* RNA-seq assays allow the direct observation of transcribed regions, splicing events, gene expression levels, rna editing events and, in some cases, transcript expression levels and transcript boundaries. The output of an RNA experiment is a set of read sequences which are then mapped back to the reference genome, producing a trace across the genome and a set of reads that only map to the genome in a gapped fashion, indicating potential splice sites (see Figure 2.3.3).

There are many variations of the RNA-seq assay, but they all share the same structure (see 2.3.4). Essentially, RNA is selected, fragmented, amplified, and sequenced. The key observation is that each RNA fragment originates from a particular transcript. Therefore, each k corresponds to a distinct transcript and θ_k is the population fraction of fragments that originated from transcript k. If we assume that RNA fragments are sampled uniformly, then every fragment of the same length is equally likely and then$\psi_{f_l}(j|l,k) = \frac{f_l}{\sum_{l' \geq l(t_k)} (l(t_k)-l) f_{l'}}$ where $l(t_k)$ is the length of transcript k. This model has been implemented in [36] and more recently in [45].

Note that because longer transcripts will, on average, produce more fragments, the $\hat{\theta}$'s are not estimates of the relative transcript frequencies. Rather, under the uniform fragmentation model, the concentration of transipt k in solution is $\frac{\theta_k}{\sum_{l' \geq l(t_k)} (l(t_k)-l) f_{l'}}$. We discuss such normalization in detail in Chapter 4.

Biases in RNA-seq read coverage have been widely reported (e.g. [5, 23]). Although refinements in experimental techniques have substantially reduced these biases in many areas,

some, such as GC bias due to differential PCR amplification, remain and will be very difficult to eliminate via improved experimental protocols. There is a large body of literature that has developed models to remove such bias, but the core component is the same. Given two fragments, x_1 and x_2, of the same length that originated from the same transcript, one might expect one to be overrepresented. For instance, for the RNA-seq data collected by [5], if x_1 has a 60% GC content and x_2 has a 40% GC content, we would expect to observe x_1 20% more often. We extend our framework to allow such bias to be modeled by re-defining $\psi_{f_l}(j, l|k)$ as $\frac{f_l \omega(k,j,l)}{\sum_{l' \geq l(t_k)} f_{l'} \sum_{m=1}^{f_{l'}} \omega(k,m,l')}$ where $w(k, j, l)$ is a weight function.

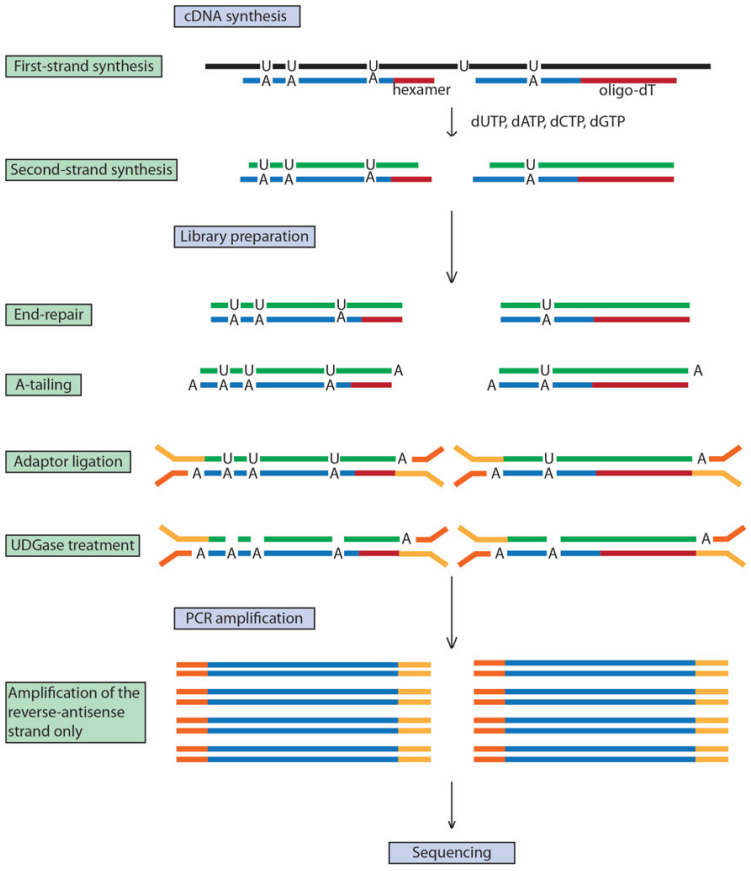

FIGURE 2.3.4. Stages in a typical RNA seq experiment.

2.3.3.2. *ChIP-seq kernel* . ChIP-seq assays allow the identification of genomic locations occuppied by DNA binding proteins. In ENCODE, it has been applied to over 2000 transcription factors and histone modifications across hundreds of different cell lines and conditions. The assay is straightforward at a high level: all proteins in a cell are attached tightly to their bound DNA via UV or formaldehyde cross-linking. The DNA is then fragmented, and a bead-attached antibody is added which binds to the DNA-binding protein of interest. The

beads are purified (with the protein and DNA still attached), the protein and DNA are unlinked, and these DNA fragments are sequenced (see 2.3.5).

The goal of a CHiP-seq experiment is to identify binding sites. The fragment corresponding to each sequenced read is known to have been bound by a protein, but from a single read it is impossible to determine exactly where in the fragment the protein was bound. By aligning multiple reads and noting that DNA can not fragment at locations bound by a protein, it is possible to more precisely identify binding sites (see 2.3.6). Note that a ChIP-seq experiment does not allow the identification of the total amount of bound DNA, but rather the ratio of the occupancies of various binding sites.

This suggests a natural form for the ChIP-Seq kernel, where each position k in the genome corresponds to a potential binding site. θ_k then corresponds to the expected fraction of fragments that originated due to a protein bound at binding site k. By assay design, such fragments are emitted with density $\psi_{f_l}(j, l|k) = \frac{f_l}{\sum_{l'} f_{l'}(l'-b_l)}$, where b_l is the length of the protein of interest's binding site.

2.3.4. Estimating θ. We use the EM algorithm and linear programming to find the maximum likelihood estimate of θ. In the M step, with θ_{old} and thus $P_\theta[g_{j,l}]$ assumed known, we calculate

$$P_{new}[g_{j,l} \,|\, r_i] \;=\; r_{ij} \sum_k \theta_k^{old} \psi_{f_l}(j|l,k) \tag{2.3.2}$$

where $\psi_{f_l}(j|l,k)$ is the fraction of fragments with length l originating from location j given that they originated due to biological event k. We then normalize $P_{new}[g_{j,l} \,|\, r_i]$ such that

$$\sum_{all\ j} P_{new}[g_{j,l} \,|\, r_i] \;=\; 1 \tag{2.3.3}$$

Then, in the E step, given the updated estimates of $P\,[g_{j,l} \,|\, r_i]$ we update the expectation of the marginal read density as

$$P_{new}[g_{j,l}] \;=\; \sum_{all\ i} P_{new}[g_{j,l} \,|\, r_i] \tag{2.3.4}$$

where $\hat{P}[r]$ is the binned frequency of r. Now we can plug back into (2.2.2) with $P_{new}[g_{j,l}]$ given by (**??**) and optimize over θ, a linear programming problem in the general case. However, in practice, the θ_k can often be estimated directly from the $P_{new}[g_{j,l}]$. This is essentially the approach of Vardi [56].

2.3.5. Confidence Bounds. Although the likelihood is convex, it is not strictly convex. Thus, although the value of maximum likelihood is guaranteed to be the maximum achievable value, there may be multiple maximums and, worse, these may be far apart in parameter space. For instance, several strains of Drosophila Melanogaster, a diploid organism, were specifically bred to reduce variation between the paternal and maternal chromosomes. Therefore, when we observe a read, it is often impossible to determine if it originated from the maternal or paternal chromosome. A similar identifiability problem is observed in retroviral insertions in the human genome. In such situations, the resulting relative posterior probabilities are completely determined by the EM algorithm's initial conditions.

For many studies, the relative read concentrations in regions with identical sequence is not important. For instance, if one is interested in estimating gene expression from an RNAseq experiment, then it is not important where the gene expression originated, just that it is being expressed. By default, we use a uniform starting location for the EM algorithm, which distributes reads equally between regions with identical sequence.

In other studies, such as a GWAS study where one is attempting to associate snps with disease traits, it is important to know where the gene expression originated. To find a lower bound for θ_k, we wish to find the lowest value that it can achieve while still being "reasonably likely". Formally, for the maxiumum likelihood estimate $\theta^\star$, we wish to find θ such that θ_k is at a minimum subject to the constraint $lhd[\theta^\star;\ \vec{r}, \vec{\psi}] - lhd[\theta^\star;\ \vec{r}, \vec{\psi}] > C$, for some choice of C. This is approximately χ^2_1 with one degree of freedom and so we choose $C = -\frac{1}{2}\chi^2_1(\alpha)$ for some desired marginal significance level α. This problem is convex, and can be solved using standard convex optimzation tools. However, it must be solved individually for each distinct k and so can be expensive in practice. We discuss this problem in detail as applied to RNA-seq data in Chapter 4.

2.3.6. Read Mapping Estimation. Given our estimate of $\vec{\theta}$, we can estimate the posterior probability that a read originated from a fragment of length l that originated from location j by

$$P[g_{j,l} \,|\, r_i] \;=\; \frac{P[r_i \,|\, g_{j,l}]\, P_{\hat{\theta}}[g_{j,l}]}{\sum_{k,l} P[r_i \,|\, g'_{k,l}]\, P_{\hat{\theta}}[g'_{k,l}]} \tag{2.3.5}$$

where only alignments with a high posterior probability are reported (by default we do not report mappings with $P[g_{j,l} \mid r_i] < 0.01$).

2.4. Approximations and Implementation Details

Maximimizing the joint likelihood naively is not feasible, and so we split the process into two stages. In the first, we identify all likely mapping locations, which we will refer to as candidate mappings. In the second, we use this sparse set of mappings to optimize the model parameters.

2.4.1. Candidate Mapping Identification. Our estimation procedure relies upon estimating

$$P[g_{j,l} \mid r_i] = \frac{P[r_i \mid g_{j,l}]\, P_{\hat{\theta}}[g_{j,l}]}{\sum_{k,l} P[r_i \mid g'_{k,l}]\, P_{\hat{\theta}}[g'_{k,l}]}$$

for each i where $P[r_i \mid g_{j,l}]$ is assumed to be known. For a sequencing experiment with 100 million reads performed on human cells using the sequencing error model described previously, this would require on the order of 10^{22} computations at each iteration of the EM algorithm. Since this is not computationally feasible, we must make approximations.

The first approximation is made by noting that, for the vast majority of k, $P[r_i \mid g_{k,l}]$ is very small, so

$$\begin{aligned} P[g_j \mid r_i] &= \frac{P[r_i \mid g_{j,l}]\, P_\theta[g_{j,l}]}{\sum_k P[r_i \mid g'_{k,l}]\, P[g'_{k,l}]} \\ &\approx \frac{P[r_i \mid g_{j,l}] P_\theta[g_{j,l}] \mathbb{I}\left[P[r_i \mid g_{j,l}] P_\theta[g_{j,l}] > \beta_1\right]}{\sum_k P[r_i \mid g_{k,l}] P_\theta[g_{k,l}] \mathbb{I}\left[P[r_i \mid g_{k,l}] P_\theta[g_{k,l}] > \beta_1\right]} \end{aligned} \tag{2.4.1}$$

where β_1 is a tuning parameter controlling the quality of our approximation. Then we need only calculate $P[g_{j,l} \mid r_i]$ for a subset of locations, although choosing β_1 to guarantee that the approximation is good may make the search very expensive in practice.

The second approximation is made by noting that given the location j with the highest value of $P[r_i \mid g_{j,l}]\, P_\theta[g_{j,l}]$, we can limit our error by ignoring locations where the ratio

$$\frac{P[r_i \mid g_j]\, P_{\hat{\theta}}[g_j]}{\max_k \{P[r_i \mid g_k]\, P[g_k]\}} < \beta_2$$

where β_2 is a second tuning parameter controlling the quality of our approximation.

The third apprximation is made by noting that the marginal read density $P_\theta[g_{j,l}]$ is positively correlated with the number of reads that map near a location j. Therefore, identifying all j

such that

$$\begin{aligned} P[r_i \mid g_{j,l}] &> \beta_1 \\ \frac{P[r_i \mid g_{j,l}]}{\max_k \{P[r_i \mid g_{k,l}]\}} &> \beta_2 \end{aligned}$$

and then plugging them back into 2.4.1 will, on average, yield a better approximation to 2.3.5. The set of genomic locations that meet these criteria are the set of candidate mappings for read i.

2.4.1.1. *Tuning Parameters.* The values of β_1 and β_2 control the tradeoff between the quality of the approximation and the computational cost. Unfortunately the scale of these parameters can vary widely between experiments, sequencing technologies, assays and even reads, which makes choosing good values difficult. To mitigate this problem, we re-parameterize $\beta_1$and β_2 in terms of a new parameter, α, defined as the expected fraction of reads at which it is acceptable for the true mapping origin to not be included in the candidate mapping set given that the read originated from the reference genome. Then we estimate values of β's to satisfy this constraint under the model.

Tuning β_1. For a desired maximum false discovery rateα and given a read r_i that originated from location g_t, we wish to choose β_1 such that $P[P(r_i|g_t) < \beta_1] < \alpha$. Under the sequencing error model, we estimate the probability of a mutation at base i to base o as $\epsilon_{i,o}$. The log likelihood is thus the product of independent multinomial with $\vec{p} = [\epsilon_{i,A}, \epsilon_{i,C}, \epsilon_{i,G}, \epsilon_{i,T}]$ and $N = 1$. In principle, we could sample from this distribution, calculate the log likelihood for each bootstrap sample, and use the empirical quantile as our estimate of $P[P(r_i|g_t) < \beta_1]$. However, this infeasible in practice, so we calculate the first two central moments and then compare to a reference distribution with matched moments.

That is, for each read we estimate the first two central moments

$$\begin{aligned} E[\log P(r_i|g_t)] &= \sum_m \sum_{o \in ACGT} \log_{10}(\epsilon_{m,o}) \\ VAR[\log P(r_i|g_t)] &= \sum_m \sum_{o_1 \in ACGT} \left[\epsilon_{m,o_1}(1 - \epsilon_{m,o_1}) + \sum_{o_2 \neq o_1} \log_{10}(-\epsilon_{m,o_1}\epsilon_{m,o_2}) \right] \\ &\quad - E[\log P(r_i|g_t)] \end{aligned}$$

then set β_1 to the α'th quartile of a gamma distribution with mean $E[\log P(r_i|g_t)]$ and variance $Var[\log P(r_i|g_t)]$. Finally, we subtract $\log_{10} 0.01$, the penalty for a single high quality mismatch, from β_1 to correct for the the fact that the log likelihood is dicrete.

Tuning β_2. We choose β_2 to limit the error of 2.4.1. Given a set of identified genomic locations $\mathbb{J}$, the total probability mass recovered is

$$\begin{aligned} &= \frac{\sum_{j\in\mathbb{J}} P[r_i \,|\, g_j]}{\sum_k P[r_i \,|\, g_k]} \\ &\leq \frac{\sum_{j\in\mathbb{J}} P[r_i \,|\, g_j]}{\sum_{k\in\mathbb{J}} (P[r_i \,|\, g_k]) + \beta_2 N G_l} \end{aligned}$$

so we set $\beta_2 = -\frac{1}{\alpha}\log_{10} G_l$, where G_l is the total number genomic locations.

2.4.1.2. *Indexing Method.* We have developed and implemented a fast indexing method specialized to work with sequencing data. It is a metric tree [55] consisting of two node types: internal nodes and leaf nodes. Internal nodes contain an array of 4^4pointers corresponding to the possible distinct combinations of AGCT. Each pointer points to another internal node or, if there are a sufficiently small number of children, a leaf nodes. Leaf nodes contain a suffix array with pointers to a union that contains either (a) a genomic location if the sequence is unique in the reference genome or (b) a pointer to an array of genomic locations if the sequence is degenerate. Since the majority of distinct sequences over 16 bases in animal genomes arise from a single genomic location, this structure helps limit the space of the genome index.

To probe the index for a specific read, we first must define a distance metric between the read and every position in the reference genome. We implement this via a length 4 array for each base in the read, where each entry in the array corresponds to the penalty of a particular genome base. For instance, under the edit distance metric, the penalty array for an A is $[0, -1, -1, -1]$. Under the sequencing error model, the penalty array for an A in position i is $\log_{10} P[r_{i1} = \mathrm{A}|\mathrm{A}]$.

Then, the search algorithm proceeds as follows:

Algorithm 1 Index Search

(1) Set the mininum allowable penalty, Ω, $\mathrm{tolog}_{10}\,\beta_1$.
(2) Initialize a queue, and add the root node of the index to it.
(3) Until the stack is empty, pop the node off the top. We will refer to this as the current node.
  (a) If the current node is a leaf node, save each genomic location where the penalty is $\leq \Omega$, and update Ω equal to the max of Ω and $penalty - \log_{10}\beta_2$.
  (b) Calculate the penalty for each letter in the current node, and add each child to the stack if the calculated penalty is greater than Ω.

2.4.2. Genome Segmentation. Naively, one would find the maximum likelihood estimate via the EM algorithm by repeated application of 2.3.4 and 2.3.2. This involves a calculation for each θ_k and every read at each step. In practice, the θ_k converge at different rates making many of the update operations unneccessary. By only updating the parameters that have not yet converged, we can significantly reduce the execution time. We accomplish this by noting that the likelihood can often be factored into a product of indepdendent components. An example of this is two transcripts that only contain unique sequence and do not overlap the same genomic sequence. In such cases, we can perform the optimization procedure over each component independently allowing us to focus on the parameters whose estimates have not yet converged.

Our goal is to factor the joint likelihood, which we formulate as a graph partitioning problem. Under this model every node corresponds to a single biological event k with $\theta_k > 0$. Nodes are connected if either (a) they have a non-zero probability of emitting a fragment that covers the same genomic base or (b) a read exists that maps to fragments that both of them have a non-zero probability of emitting. Then, the sets of connected reads and θ's correspond to the products in the factored likelihood. Note that our ability to factor the likelihood relies upon a sparse candidate mappings set; if $P[r_i|g_{j,l}]$ were never zero then the graph would be fully connected. Even after this segmentation, low quality reads can produce weak connections between otherwise distinct clusters. Thus we separate component pairs connected only by shared mappings if the ratio of the sum of the edge weights connecting them to their total mass is less than a third tuning parameter β_3. We set β_3 to $1e^{-6}$ by default.

2.5. Validation

2.5.1. Peak Calling in a Diploid Background. One significant advantage of Statmap is its ability to distinguish between long regions of very similar sequence. This, for instance, allows us to call peaks in recently diverged paralogs or differentiate between chip peaks in a diploid background. It is well known that mapping strategies dependent on “uniquely" mapping reads fail in these and similar cases.

To examine this advantage, we simulated a ChIP-seq experiment in a synthetic genome composed of two 5000 basepair 'chromosomes'. The first (paternal) is a copy of the eve stripe 2 locus taken from D. Mel; the second (maternal) is a duplicate with three single basepair mutations (simulating three single nucleotide mutations). Next, we sampled 1000 35 basepair single-end reads from the region, mutated the sampled sequences using the Illumina error model, and mapped them back to the synthetic genome. The true density is plotted, below, against the estimated densities from Statmap and Bowtie [**?**]. The bootstrap

samples were taken from 25 random local minimums with 25 bootstrap samples from each, generated using the method outlined in Section C.2.1(c). The simulation code is distributed with Statmap. All of Statmap's tuning parameters were set to their defaults with version Statmap 0.2.0 Beta 3. We used Bowtie version 0.12.5 with the –all and –try-hard flags.

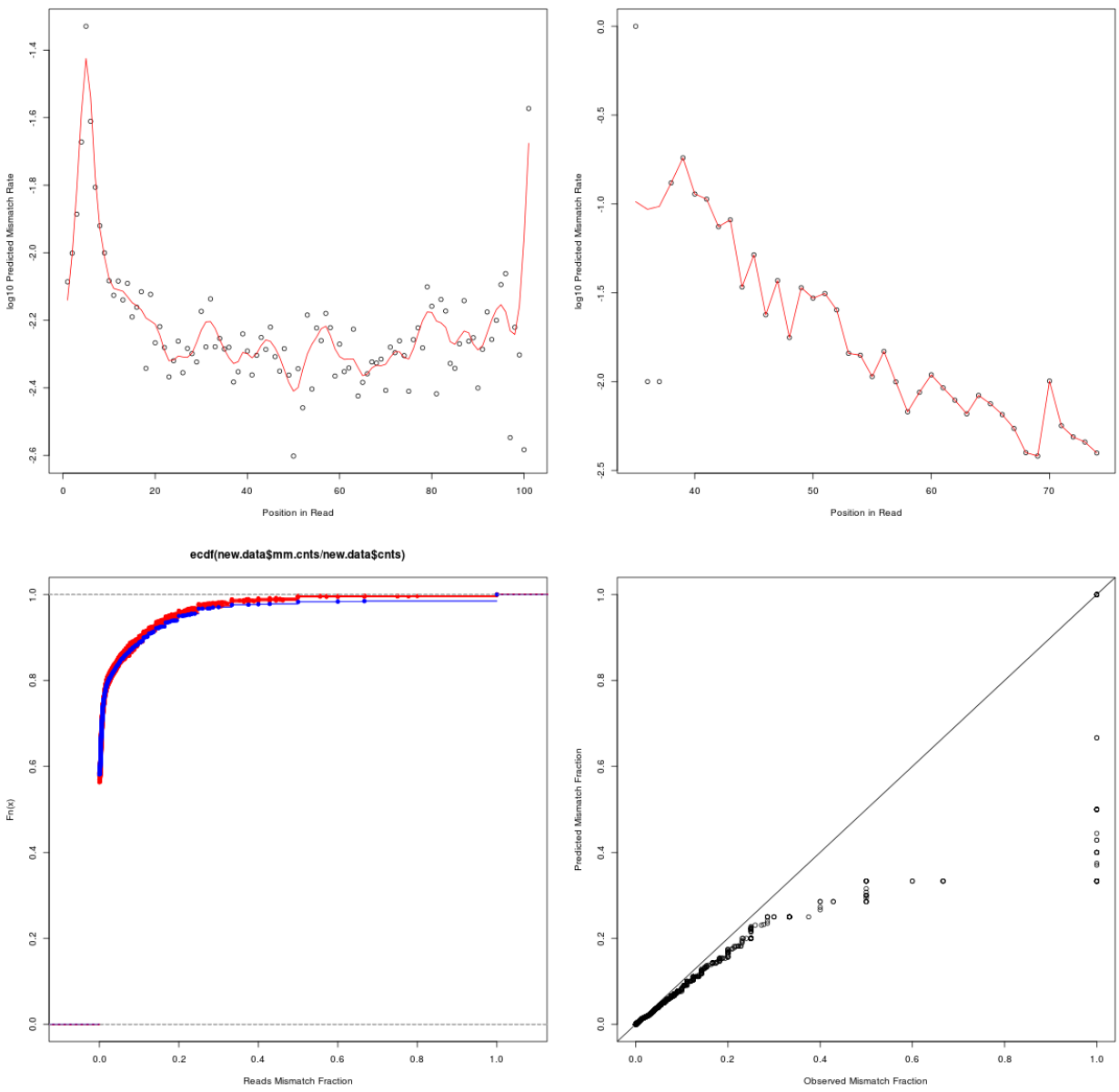

FIGURE 2.3.2. Example estimated error rates from an ATAC-seq experiment performed on 4-8 hour old *C. Elegans* embryos. The estimates correspond to pair1 reads that mapped to forward genome. The upper left and right panels show the estimated error rates versus read position and sequencing machine error estimate respectively; the red line corresponds to the smoothed estimate. The bottom left panel shows the cumulative distribution of average mutation rates by read. The blue line corresponds to the observed rates; the red lines are 20 sets of reads simulated under the fit error model. The bottom right plot shows the expected error quantiles versus observed. Although the independence model underestimates the true error rate in the tails, >99.9% of reads have estimated error rates that are within the expected sampling variance.

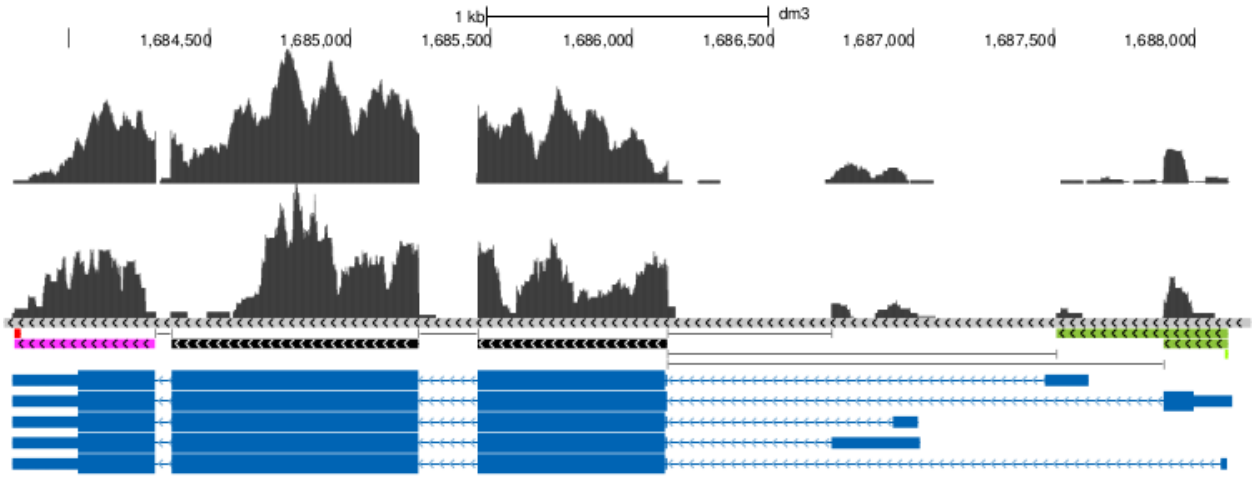

FIGURE 2.3.3. RNA-seq read coverage and identified junctions for CG2017 in two biological replicates of *drosophila melanogaster* heads from mated 20 day old females. The top two grey tracks show the mapped RNA-seq read coverage. Below are identified elements: dark green are TSS exons, black are internal exons, purple are TES exons, and thin lines are splice junctions. The bottom track contains all transcript isoforms annotated in Flybase version 12.

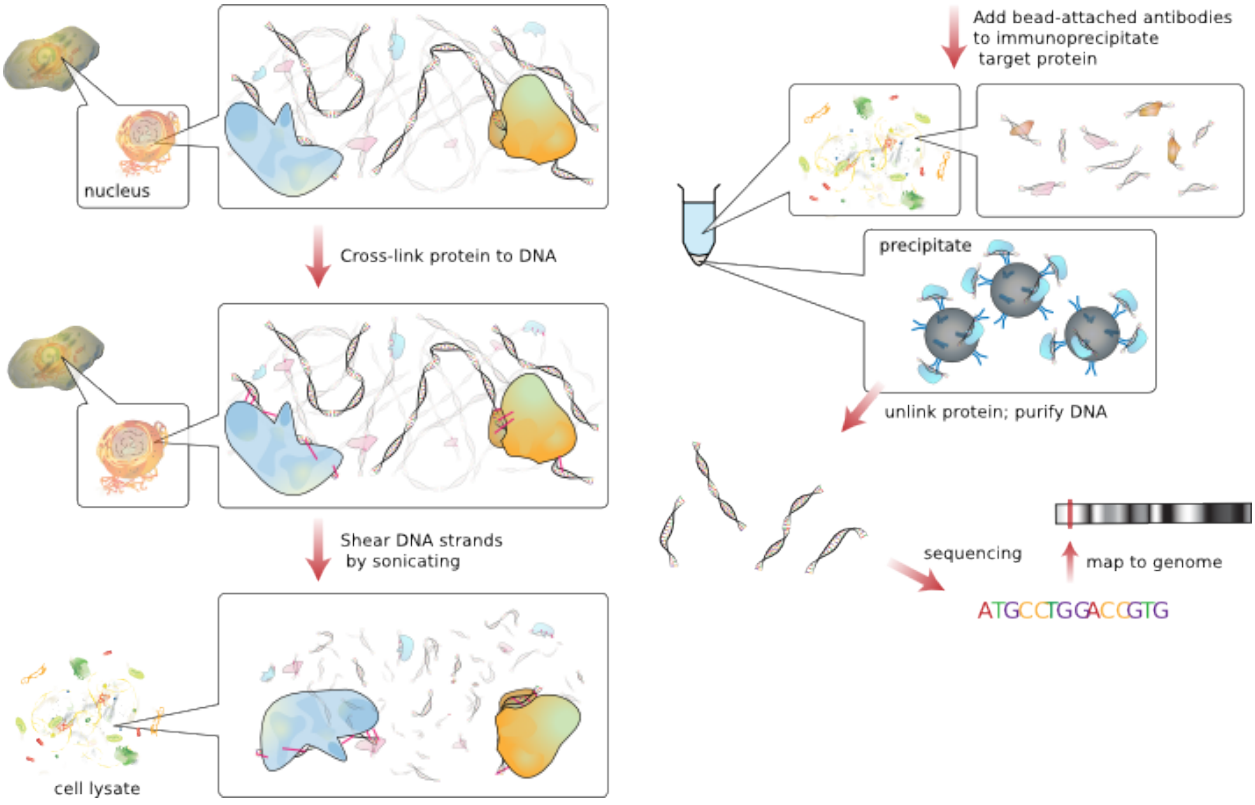

FIGURE 2.3.5. Stages in a typical ChIP-Seq experiment. Figure adapted from "Chromatin immunoprecipitation sequencing" by Jkwchui, which is licensed under Creative Commons Attribution-Share Alike 3.0 via Wikimedia Commons.

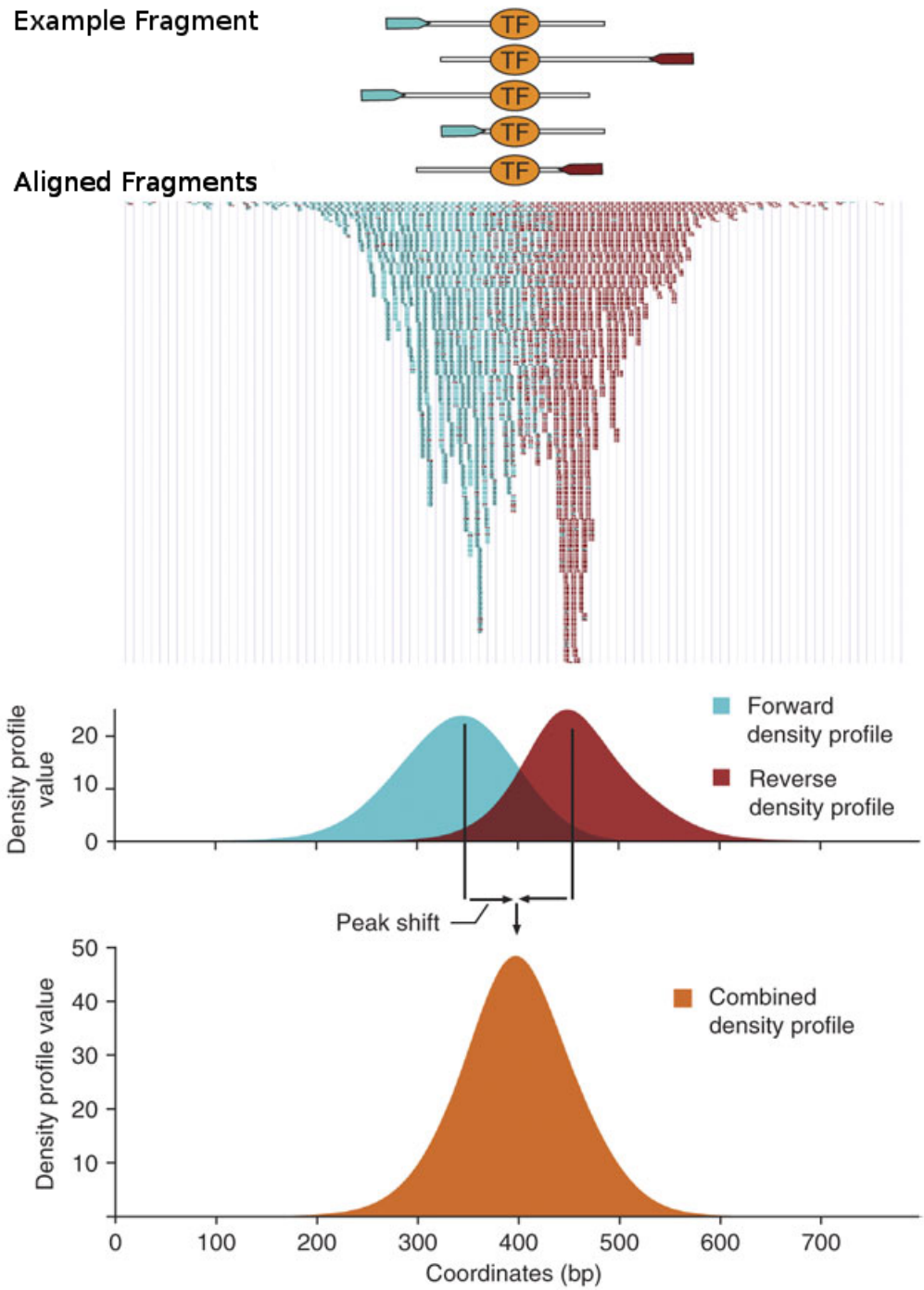

FIGURE 2.3.6. Aligned reads from a ChIP-Seq experiment. The transcription factor (TF) is bound on each fragment, but the positioning is not consistent. When the fragments are aligned then the binding site can be seen to be between bases 375 and 425. Figure adapted from http://bioinformatics.cineca.it/cast/.

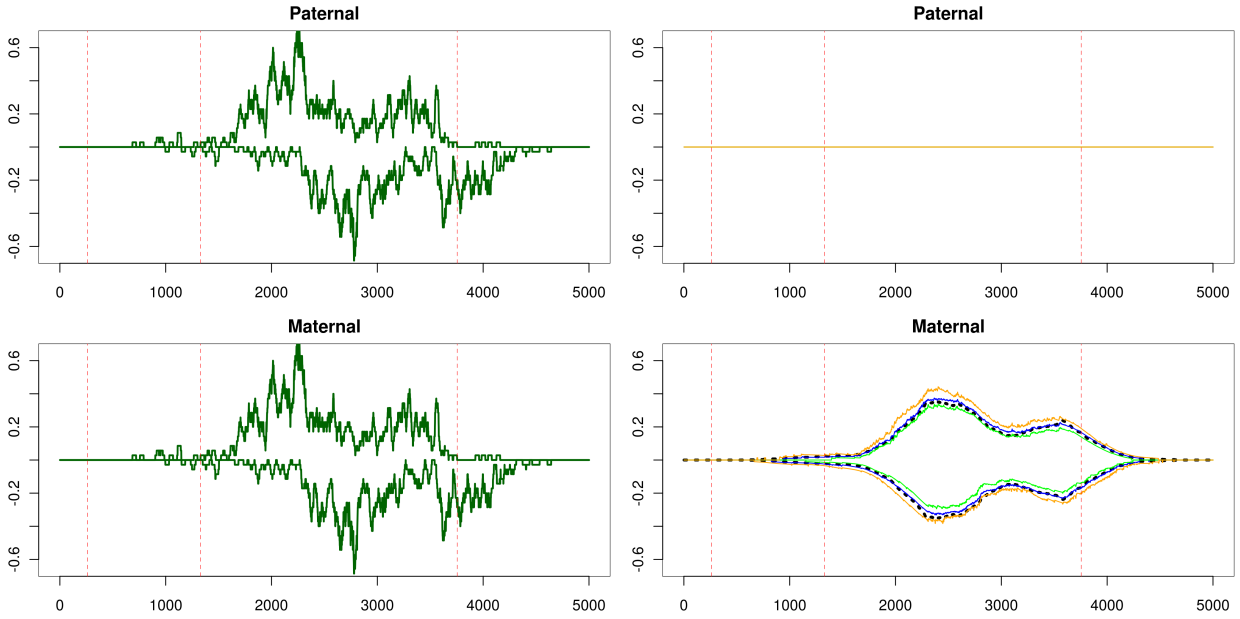

FIGURE 2.5.1. A ChIP-seq experiment was simulated in a synthetic genome composed of two 5000 basepair 'chromosomes'. The first (paternal) is a copy of the eve stripe 2 locus taken from D. Mel; the second (maternal) is a duplicate with three single basepair mutations (simulating three single nucleotide mutations). Next, we sampled 1000 35 basepair single-end reads from the region, mutated the sampled sequences using the Illumina error model, and mapped them back to the synthetic genome. The left plots correspond to the bowtie estimated read coverage; the right to stamps. The true density is plotted in a dashed black line, blue is the statmap estimated density, i.e. the maximum likelihood estimate of 2.2.2. Orange and green are the upper and lower bounds, respectively.

CHAPTER 3

Identifying Transcript Bounds

3.1. Overview

Genome regions are transcribed from DNA into RNA, and then "processed". During processing, contiguous segments of RNA are removed, a string of A's (i.e. poly(A) tail) is added to the end of the transcript, and a 'cap' is added to the beginning. Processed transcripts are transcripts after these processing steps have occurred.

There are three primary experiments for interrogating transcript structure. RNAseq provides a simple random sample from all transcript fragments, primarily providing information about the internal transcript structure. Poly(A)-site-seq sequencing (PASseq) targets sequences with poly(A) tails, and thus provides data about the location of processed transcript end sites (TESs). The CAGE assay, and it's successor RAMPAGE, sequences capped fragments, providing a sample from transcript start sites (TSSs). We focus our analysis efforts on the latter two assays, but RNAseq data can serve as a useful control. All three assays are noisy. For instance, the cap selection step in a CAGE assay is not 100% specific, so the actual distribution of observed reads is a mixture of capped and uncapped transcript fragments. Peak detection is the process of separating noise from signal, or identifying regions in the genome that are significantly enriched for TSS or TES sites.

3.2. Motivation and Previous Work

The FANTOM consortium was funded to identify all human TSS's and, to this end, performed the CAGE assay in 114 distinct human tissues. The set of TSS's identified by the consortium was 10 fold larger than the sum of all previously identified TSSs. Disturbingly, the novel TSSs were often located in genomic regions previously identified as coding and 3' UTRs, which is inconsistent with our current understanding of how different sequence composition corresponds to different function. Furthermore the RIKEN-identified TSS regions correspond poorly to promoter regions identified from histone modification data (see 3.2.1). The fact that transcribed regions are enriched for CAGE tags suggested to us that the cap selection process is not specific. We have developed a model which identifies such noise in order to estimate the distribution of TSS's.

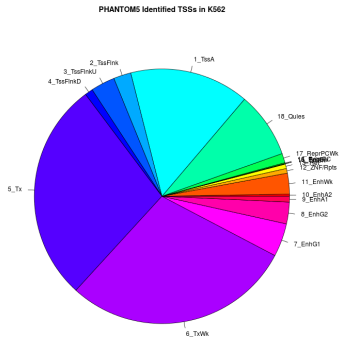

FIGURE 3.2.1. All K562 TSSs identified by the FANTOM5 consortium were associated with their chromatin state class membership as predicted by the Human Epigenome Roadmap. The majority of the TSSs fall in "transcribed" classes (purple) and the TSS classes (blue) only account for 20.4% of the total signal.

3.3. Method

For pedagogical reasons we present our method in terms of a CAGE experiment; however, the technique applies equally well to PASseq and RAMPAGE data.

3.3.1. Mixture Model. We model the CAGE data as a mixture of reads sampled from the noise distribution (uncapped fragments) and signal (capped fragments). The set of observed reads from a properly matched RNA-seq experiment (e.g., same biological sample, same nucleic acid extraction method, etc.) is essentially a sample from the uncapped, or noise, distribution. Furthermore, we expect the distribution of noise reads within an exon to be roughly uniform. We do not have any prior knowledge about the shape of the capped distribution, but we do expect there to be few bases which correspond to true TSS's and we expect these bases to be situated near each other. To formalize this, we first must define some terminology:

- Y_i the observed count of CAGE reads that start at base i
- N_i the observed count of RNAseq reads that start at base i
- ψ_i the population fraction of signal reads that begin at base i
- η_i the population fraction of noise reads that begin at base i
- λ the fraction of CAGE reads in a region that originated from the noise component

We assume that the probability of sampling a read starting at position i from the population of CAGE reads is $\lambda\eta_i + (1-\lambda)\psi_i$ and that the probability of sampling position i from the

population of RNA-Seq reads is η_i. If we do not constrain the distribution of ψ andη then the log likelihood is that for a mixture of two multinomials

$$\begin{aligned} l(n,s;Y,N) &\propto \sum_i \{Y_i \log(\lambda\eta_i + (1-\lambda)\psi_i) - \log(Y_i!)\} + \sum_i \{N_i \log(\eta_i) - \log(N_i!)\} \\ s.t. &\quad \sum_i \eta_i = 1,\ \sum \psi_i = 1,\ \psi_i \geq 0,\ \eta_i \geq 0 \end{aligned}$$

and a maximum likelihood estimate is $\lambda = 0$, $\eta_i = \frac{N_i}{\sum_{N_i}}$, and $\psi_i = \frac{Y_i}{\sum Y_i}$, although this estimate is not unique. Although suggestive, the maximum likelihood estimate does not encode any of our prior information on the distributions of $\vec{\psi}$ and $\vec{\eta}$; namely, that $\vec{\psi}$ is smooth within exonic regions and $\vec{\eta}$ is sparse both in the total number of non-zero bases (i.e., bases with $\eta > 0$) and the number of contiguous intervals with any non-zero base.

We have developed a heuristic procedure that identifies such regions, which we describe below.

3.3.2. Estimation Procedure. Our procedure has three distinct parts. In the first, we segment the genome into gene regions, which we describe in 5. In the second, we use the RNA-seq control to estimate the distribution of $\vec{\eta}$. In the third we condition on $\vec{\eta}$ and then greedily find regions that appear to be enriched for reads taken from the signal distribution.

3.3.2.1. *Estimating the Density of the Noise Distribution.* RNA-seq experiments are designed such that all fragments originating from the same transcript and have the same length are equally likely to be observed. We do not directly observe the transcript that a fragment originates from, but we do know the genomic locations corresponding to the regions that transcribed to produce the fragment. Within a contiguous genomic region with no TSS, TES, or splice sites, the expected distribution of RNAseq fragment start sites is uniform. Given all such regions we could then, in principle, estimate the noise fragment density by the average number of fragments that begin in the region.

In practice, the fragmentation process is not perfectly uniform and all TSS, TES, and splice sites are not known. Therefore, we use reference and discovered TSS, TES, and splice sites to segment a gene region. Furthermore, within each segment, we use a kernel density estimator to smooth the RNAseq counts which helps to make our estimate robust.

3.3.2.2. *Identifying Signal Regions.* Our goal is to identify the regions with $\vec{\psi}$ greater than 0. However, even if $\vec{\eta}$ were known, the model is not identifiable so we need additional constraints. Biological knowedge motivates us to prefer solutions where both the total number of bases with $\vec{\psi} > 0$ is small and the number of contiguous regions with $\vec{\psi} > 0$ is small.

Finally, since our goal is to produce a set of high confidence TSS's, we prefer solutions with high values of λ.

This problem is not necessarily well defined as there is a fundamental tradeoff in the number of identified regions and the number of called bases. To take an extreme example, if we were only concerned with minimizing the number of enriched regions then we could call the entire gene region enriched. Of course, this gene region wide "peak" would include a large number of bases that are not enriched for signal reads. We have developed a heuristic algorithm based upon a hierarchical testing approach which we believe produces peaks that are a reasonable trade-off between these two competing interests.

Our algorithm requires us to decide whether a particular sub-region is enriched for signal, which we formulate as a hypothesis testing problem. That is, for a genomic region, we wish to test the null of every $\psi_i = 0$ (i.e. $\theta_0 = \sum \psi_i$) versus the alternative of any $\psi_i > 0$ (i.e. $\sum \psi_i > 0$). If we can reject the null at level α then we declare the region enriched - otherwise we assume the region is all noise. We dicuss the details of the test in 3.3.2.3.

The algorithm depends on the mixture parameter λ which is unknown. To be conservative we initialize $\hat{\lambda} = 1$, call peaks and then update our estimate of $\hat{\lambda}$ to $\sum_{i=1}^{L} Y_i \mathbb{I}[i \in \mathbb{N}] / \left(\sum_{i=1}^{L} Y \sum_{ii=1}^{L} \eta_i \mathbb{I}[i \in \mathbb{N}] \right)$, where $\mathbb{N}$ is the set of noise regions. We repeat the peak calling process until our estimate stabilizes. Note that, because we are using regions in which we failed to reject the null at level $\alpha < 0.5$, $\hat{\lambda}$ is always an overestimate of the true noise fraction.

Peak Calling Algorithm Overview. Given a gene region of length L, an estimate of $\vec{\eta}$, a minimum region size, and a desired significance level α we first test whether the gene region is enriched at level α. If we can't reject the null, then we declare the region devoid of signal reads and are done. Otherwise we proceed to refine the region.

Note that under our model this initial test controls the type I error for the number of falsely identified enriched regions. But, because rejecting the null only tells us that at least 1 base in the region is enriched for signal reads, it tells us very little about the number of bases that are incorrectly identified as enriched. This is where the heuristic comes into play.

Since we now know that 1 base is enriched, and we expect the enrichment to be local and contiguous, we ask ourselves if we can explain the enrichment by 1 or more enriched subregions. So we randomly split the enriched into two regions, test each for enrichment, and thus have 3 outcomes:

1) we only reject the NULL in one region

In this case, we add the failure to reject region to the list of noise regions, and the enriched region to the list of regions to further refine. This always allows us to improve our estimate regardless of the desired region/base enrichment trade-off. That is, since we have the prior belief that the number of enriched bases and regions is small, we assume that the enrichment of the parent region is due to the sub-region in which we were able to reject the NULL. Now we have the same number of enriched peaks, but less enriched bases. Of course this increases the expected number of type II errors.

2) we do not reject the NULL in either region

In this case we declare the parent region enriched, and add it to the called peaks list. This is bad for our ability to reduce the number of identified signal bases but good because we do not identify any additional enriched regions.

3) we reject the NULL in both regions

In this case we add both regions to the list of regions to further refine. Since we have the power to detect enriched bases in both regions we can further refine the peaks list by reducing the number of enriched bases, but it also increases the number of called regions.

We continue with this process until we reach a point where we have a set of regions that are enriched, but we can't explain the enrichment in these regions by further subdivisions.

The precise algorithm is described in 2.

Split Method. We have presented the algorithm in context of a random split point because it allows us to control the type I error rate for every called enriched region. Running the algorithm using a random split point can yield inconsistent results. To improve our estimate, we run the algorithm multiple times and call a base enriched only if it is enriched in the majority of runs. However, this process is expensive computationally, and so we now choose the split point at the largest sub-region with zero signal reads.

Note that by looking at the data to choose our split point, we can no longer make guarantees about the type I error rate at the enriched region level. For instance, if one chose the split point to maximize the chance of rejection in the left subregion, then the type I error rate for that region would be higher than α. That being said, we see no reason why our split point method would be dependent on the distribution of the test statistic. Furthermore, empirically, our split point method gives similar results to the random split point method and so we do not believe that our split point heuristic is inflating the peak level type I error rate.

Algorithm 2 Greedy identification of singal regions

Given a gene region of length L, an estimate of $\vec{\eta}$, a minimum region size, and a desired significance level α:

(1) Initialize $\lambda_0 = 1$
(2) Initialize the set of noise regions ($\mathbb{N}$) and signal regions ($\mathbb{S}$) to the empty set
(3) Initialize the set of regions to test, $\mathbb{T}$, to contain the entire gene region
(4) Until $\lambda_i = \lambda_{i-1}$
  (a) While $\mathbb{T}$ is non-empty, choose a region from $\mathbb{T}$, and test the region for significance at level $\alpha/2L$ (see 3.3.2.3).
    - If the region is not significant, add it to $\mathbb{N}$
    - If the region is significant and it is smaller than the minimum region size, add it to $\mathbb{S}$
    - Otherwise, split the region into 2 subregions by choosing the base with the lowest number of reads and add them to $\mathbb{T}$.
  (b) Update the estimate of lambda, setting it to

$$\lambda_i \;=\; \sum_{i=1}^{L} Y_i \mathbb{I}[i \in \mathbb{N}] / \left(\sum_{i=1}^{L} Y \sum_{ii=1}^{L} \eta_i \mathbb{I}[i \in \mathbb{N}] \right)$$

3.3.2.3. *Testing for Significance of a Region.* Given a region covering bases $[i, i+n)$, we wish to determine whether $\psi_i = 0$ for every position within the region. Defining $\theta = \sum_{j=i}^{i+n} \psi_j$, this is equivalent to testing the null $\theta_0 = 0$ versus the alternative $\Theta_1 = \{\theta | \theta > 0\}$. Given$\lambda$, $\vec{\eta}$, the log-likelihood ratio is

$$\log \left[lhd(\theta_0; \vec{Y}) / lhd(\Theta_1; \vec{Y}) \right] \;=\; \sum_{j=i}^{i+n} \{ Y_j \log \left((1-\lambda)\eta_j \right) \} - \sum_{j=i}^{i+n} \{ Y_j \log \left(\lambda \eta_j + (1-\lambda)\psi_j \right) \}$$

Note, critically, note that the likelihood ratio statistic is non-decreasing in θ. Thus, were the $\vec{\psi}$ known, the universally most powerful test at level α would set the critical value C such that $P\left[lhd(\theta_0; \vec{Y}) > C \right] = \alpha$. Since the ψ_j are unknown, we choose them to maximize $lhd(\theta_1; \vec{Y})$ subject to $\theta_1 = \sum_{j=i}^{i+n} \eta_j$ and $\eta_j \geq 0$, which is also non-decreasing in θ_1, so we chosoe $P\left[lhd(\theta_0; \vec{Y}) > C \right] = \alpha$.

In principle, we can estimate the critical value with the parametric bootstrap. We sample $N_{bootstrap}$ times from the multinomial with bin probabilities $\vec{\eta}$ and counts $\lambda \left(\sum Y_i \right)$, and then estimate the critical value by the α'th empirical quantile among

$$\sum_{j=i}^{i+n} \left\{ N_j^{(k)} \log \left(\eta_j \right) - \log \left(N_j^{(k)} ! \right) \right\}$$

where $N_j^{(k)}$ is the count from base j in the k'th sample. Unfortunately, this procedure is impractical for values of α less than $1e^{-2}$, which is very common because of the necessary multiple testing corrections.

Under the null, the bin counts are nearly independent and so the distribution of counts at base i can be approximated by the binomial $Bin\left(\lambda\left(\sum Y_i\right), \eta_i\right)$. Since the η_i are typically very small, we can calculate the moments efficiently with the truncated series

$$Ef(X^m) = \sum_{k=0}^{\lambda \sum Y} \binom{\lambda \sum Y}{k} \eta_k^k (1-n_k)^{\lambda \sum Y - k} \left(k \log\left(\eta_i\right) - \log\left(k!\right)\right)^m$$

$$Ef(N_i)^m = \sum_{k=0}^{\lambda \sum_i Y_i} \frac{\eta_i^k}{k!} e^{-n_i} \left(k \log\left(\eta_i\right) - \log\left(k!\right)\right)^m$$

where we truncate the sum when the last term multiplied by $\lambda \sum Y - k$ is below some threshold, thus bounding our relative error. We then estimate the critical value by $\Phi_\eta^{-1}(\alpha)$, where Φ is a gamma distribution with matched moments.

3.4. Results

3.4.1. Comparison to GENCODE annotated TSS's. The Gingeras lab, working with the ENCODE consortium, has performed 36 RAMPAGE experiments and matching RNAseq data. All of the experiments were performed on Ribosome depleted, >200 basepair RNA fragments collected from whole cells. In addition, the Gingeras lab produced peak calls on each of these exeriments. Here we compare the peaks identified by our method to those produced by the Gingeras lab in two samples: neural embryo tissue and the K562 cell line.

We mapped the RAMPAGE and RNAseq reads to the hg19 reference using the ENCODE consortium's mapping pipelines. In addition, for the fetal neural sample, we eliminated PCR duplicates by removing RAMPAGE reads that had identical sequence to another reads. We ran the peak calling algrotihms on these reads and compared the resulting peaks to GENCODE v19. We identified a peak as a match if it was within 50 basepairs of an annotated GENCODE TSS, and plotted the fraction of GENCODE TSSs identified versus the fraction of peaks that were also identified in GENCODE. Of the top thousand peaks called by both methods, 93.1% of ours were within 50 basepairs of a GENCODE annotated TSS versus 78.7% for the Gingeras lab's method. However, in the top 10,000 peaks called by each method, 84.6% of ours correspond to a GENCODE TSS versus 83.0% for the the Gingeras lab method. See 3.4.1 for full the specificity versus sensitivity plot.

We performed the same analysis on RAMPAGE data collected from the K562 cell line. In addition, we applied our method to K562 CAGE data collected by the FANTOM5 consortium, and their identified peaks. Again, if the GENCODE TSS's are taken as the truth, then our

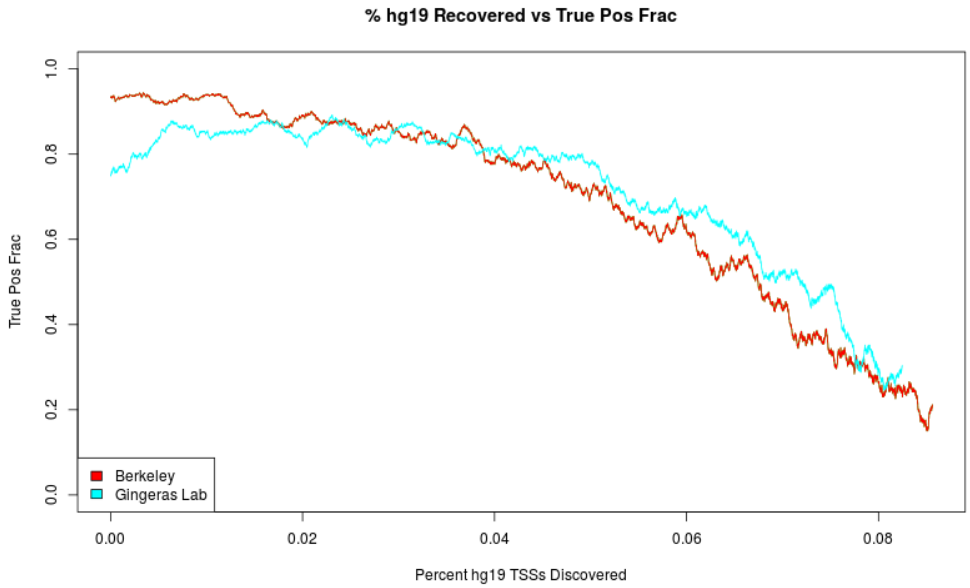

FIGURE 3.4.1. Fraction of GENCODE version 19 recovered versus average recovery rate. Note that our method produces less false positives among the high quality peaks. The majority of these peaks appear in heavily transcribed regions, which we are able to flag as false positives due of our use of an RNAseq control.

method produces consistently higher quality peaks than other methods (see). Furthermore, the quality of called peaks is similar between biological replicates and even across assays.

3.4.2. Comparison to Human Epigenome Predicted States. Comparing peak calls to GENCODE annotated TSSs is useful, but might give us a biased estimate of the accuracy of our peak-calling method. The Human Epigenome Roadmap Project collected histone modification data across hundreds of tissues and cell types, and used this data to classify genomic regions in each sample, including K562. When we analyze the chromatin predicted states of each TSS predicted by the various methods, we find that our method has the highest enrichment of the TSS and TSS flanking classes. Although it appears that the CAGE data is significantly worse than the RAMPAGE, the performance gap may be exaggerated by the lack of a matching RNAseq sample from the same cell growth batch.

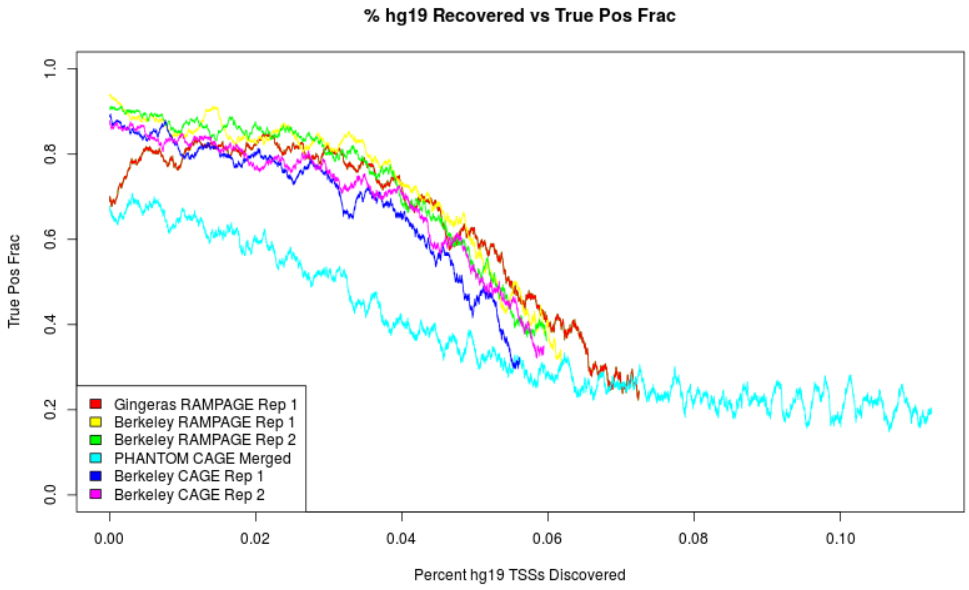

FIGURE 3.4.2. Fraction of GENCODE version 19 recovered versus average recovery rate. Note that, although FANTOM CAGE peaks correspond very poorly with GENCODE annotated TSSs, by integrating a matching RNAseq data set we are able to produce a high quality peak list from the same input data.

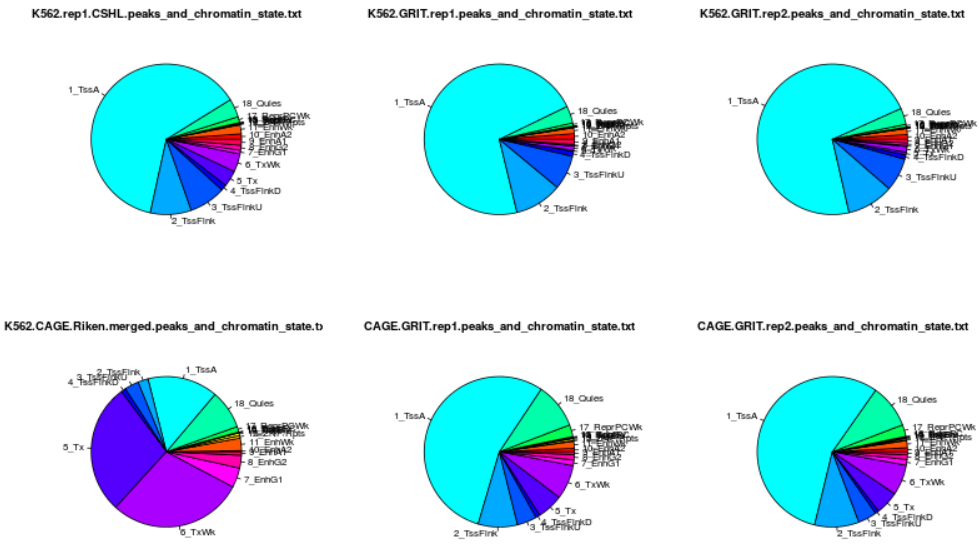

FIGURE 3.4.3. Chromatin state enrichment of predicted TSSs.

CHAPTER 4

Transcript Expression Estimation

4.1. Introduction

A typical RNA-seq experiment involves by purifying an RNA sample, fragmenting the RNA, converting these fragments to their equivalent cDNA with reverse-transcriptase, amplifying the cDNA via PCR, and finally sequencing the resulting fragments. The resulting sequences are then mapped back to a reference genome. For clarity, we will assume that the mapping procedure is perfect, i.e. we assume that we can unambiguously identify the genomic region from which each fragment originated.

One of the primary uses of RNA-seq data is to estimate the concentration of various gene elements within a sample. For instance, a researcher may wish to identify genes that are over-expressed in a cancer sample in the hope of identifying onco-genes or potential drug targets. The primary challenge in estimating gene expression is relating the number of reads that map to a gene to the number of RNA molecules in the sample. The fragmentation process is designed such that each fragment of a particular type is equally likely, and so we apply the following generative model:

Algorithm 3 Generative RNAseq model

(1) Choose a fragment length with probability f_l
(2) Choose a gene of length L with probability $(L_{g_i} - f_l)g_i / \sum_k (L_{g_k} - f_l)g_k$
(3) Choose position j with probability $1/(L_{g_i} - f_l)$, which yields the fragment spanning bases $[j, j + L - 1]$

Limitations in sequencing technologies limit the length of the fragments that can be efficiently sequenced to roughly 400 basepairs, much shorter than a typical human mRNA. In this limit, $L_g \ggg f_l$, a gene's relative concentration of gene i is just $\frac{N_{g_i}/L_{g_i}}{\sum_k N_{g_k}/L_k}$. A commonly used measure of gene expression is the number of sequenced fragments that map to a gene per kilobase of gene length per million reads sequenced (FPKM) which is only correct in this limit. We can improve the estimate by accounting for the fragment length distribution. Given the relative concentration of a fragment of length l, f_l, we estimate the relative concentration of gene i by $\sum_l f_l \frac{\mathbb{I}(L_{g_i} \leq l)\left(N_{g_i} - l\right)}{\sum_k \mathbb{I}(L_{g_k} \leq l)\left(N_{g_k} - l\right)}$ where$\sum_l f_l = 1$. This model is used by the most popular gene expression estimation tools (robinson2010edger, delhomme2012easyrnaseq, anders2012differential)

although these tools also correct for additional assay biases and mappability. However, even after accounting for such biases, this model is wrong because fragments arise from transcripts and genes may produce multiple transcripts of varying lengths. The obvious correction is to estimate expression at the transcript level, but this leads to a different set of problems.

4.2. Transcript Expression Estimation

4.2.1. Frequency Estimation. The primary challenge in estimating transcript expression is identifying a vector, $\vec{t}$, that corresponds to the transcripts relative concentrations in solution. This is difficult because reads can not necessarily be unambiguously assigned to one transcript. Therefore, the first step in estimating transcript expression levels is redefining the transcripts in terms of non-overlapping exon segments, or pseudo exons. It is then possible to unambiguously group reads by the set of pseudo exons that they overlap, which we refer to as a "fragment type" (see Figure 4.2.1). The fragment types that can be directly observed is a function of both an RNA-seq experiment's sequenced read length and fragment length distribution.

The fraction of reads that are expected to be of a certain fragment type is used to build the design matrix, X. Formally, each entry X_{ij} is defined to be the probability of sampling fragment type i given that the read originated from transcript j. In practice, we estimate X_{ij} by $\frac{\sum_l \hat{f}_l C^l_{i,j}}{\sum_{k=1}^{N_t} \sum_l \hat{f}_l C^l_{k,j}}$ where $\hat{f}_l$ is the estimated fraction of fragments of length l, $C^l_{i,j}$ is the count of distinct fragments of length l in transcript j that produce fragments of type i, and N_t is the total number of transcript models. This estimate formalizes the assumption that, within a transcript, all fragments with the same length are equally likely to be observed.

Then, given a vector of observed bin counts, $\vec{Y}$, the maximum likelihood estimate of the transcript frequencies, $\vec{t}$, is the vector $\hat{t}$ that maximizes the log-likelihood, $lhd(Y; \vec{t}) = \sum_i Y_i \log\left[X\vec{t}\right]$, subject to the constaints that $t_j \geq 0$ and $\sum_j t_j = 1$. This estimate is unique whenever no row of X can be constructed from a positively weighted sum of other rows. In such unique cases, the statistical model is said to be identifiable given the data.

Maximizing the likelihood equation requires optimizing $lhd(Y; \vec{t}) = \sum_i Y_i \log\left\{\sum_j X_{ij} t_j\right\}$, subject to $\sum_j t_j = 1$, $t_j \geq 0$. Although this is convex and can be solved using standard convex solvers like CVX [20], the potentially large number of candidate transcripts makes such approaches too expensive to use routinely. We have found that, in practice, a projected gradient ascent method is the most performant (data not shown). We find a starting location by minimizing $\sum_i \left(\frac{Y_i}{\sum_j Y_j} - \sum_j X_{ij} t_j\right)^2$ st $\sum_j t_j = 1$, $t_j \geq 0$ using a QP solver. Then, we use projected gradient ascent with a fast simplex projection method [14] until the update

differences are less than machine precision. Since the likelihood surface is smooth and convex, this method always converges to the optimum. We have verified that solutions found by the GRIT software package are equivalent to the CVX solutions (data not shown).

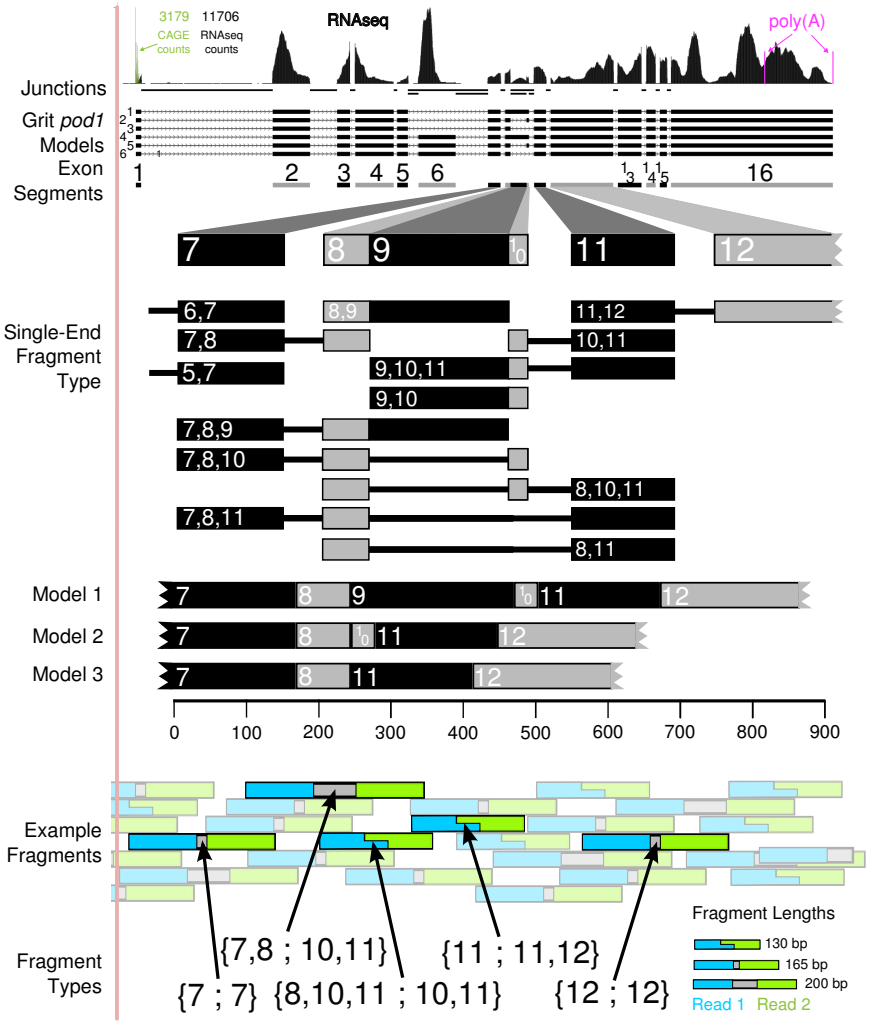

FIGURE 4.2.1. **Expression Estimation Overview** - To identify the set of transcripts in *pod1,* we find the set of non-overlapping segments, labeled **exon segments**, with which it possible to reconstruct the transcript set. In the zoomed-in region containing segments 7-12, the possible fragment types, labeled **Single-End Fragment Types**, that can be observed from 75 basepair reads are shown. Next, we estimate the fragment length distribution, and then identify the sets of pseudo exons that can be overlapped by paired end reads. The blue and green fragments are possible fragments taken from transcript model 2. For example, in the 200 basepair fragment labeled {7,8;10,11}, read 1 (in blue) overlaps exon segments 7 and 8, while pair 2 (in green) overlaps segments 10 and 11. The fact that read 1 overlaps segments 7 and 8 doesn't give us any additional information about the transcript isoform from which it originated, but the fact that read 2 overlaps 10 and 11 implies that it must have come from either model 2 or 5.

4.2.2. Confidence Bounds. To form confidence bounds on a particular transcript's frequency estimate, $\hat{t}_i$, our goal is to find the minimum and maximum values that t_i can take while still being "reasonably likely" to produce the observed data set. We identify a subset Δ_R of the probability simplex such that $lhd(Y; \vec{t})$ is sufficiently high for every $\vec{t} \in \Delta_R$. Convexity of the likelihood function guarantees this region is simple and convex, which allows us to form our confidence bound for transcript i as the interval $[\min\{t_i : t \in \Delta_R),\ \max\{t_i : t \in \Delta_R)]$ - a conservative estimate for individual coverage rates.

This interval can be estimated directly by finding the $\vec{t}$ on the probability simplex that minimizes t_i such that the log likelihood ratio $lhd(\vec{t}_{mle}) - lhd(\vec{t})$ is above some critical value. Formally, we t_i, involves finding the minimum value of $\vec{t}$ which minimizes the i'th component, subject to the restriction that the log likelihood ratio $lhd(\vec{t}_{mle}) - lhd(\vec{t})$ is sufficiently high. We use the objective $lhd(Y; \vec{t}) = Y_O^{t_O} + \sum_i Y_i \log\left\{\sum_j X_{ij} t_j\right\}$ where t_O and Y_O are the estimated fraction and the count of reads that fall outside the the gene of interest. This objective accounts for the fact that the number of reads that originates from a given gene locus is random. Because the maximum likelihood estimate of t_O is $\frac{Y_O}{Y_o + \sum_i Y_i}$, we rescale $\vec{t}_{mle}$ by $1 - \hat{t_O}$ to calculate $lhd(\vec{t}_{mle})$.

Since the asymptotic distribution of $lhd(\vec{t}_{mle}) - lhd(\vec{t})$, a log likelihood ratio statistic[3] with one degree of freedom, is $\frac{1}{2}\chi^2$ we set the critical value to $\frac{1}{2}\chi^2(\alpha)$ for some desired marginal significance level α. When the model is identifiable, simulations show that this approach produces confidence bounds with the correct rejection rates for realistic sample sizes (see Figure 4.2.2). For unidentifiable models, our method produces a lower confidence bound of zero for every transcript in the gene. This allows the user to easily identify regions in which RNA-seq data alone is not sufficient to identify the set of transcripts present. In contrast, Cufflinks and Rsem[35] both use a Bayesian approach, sampling from a posterior distribution to estimate confidence bounds. In complex genes, such as Dscam1 or Mhc, the resulting confidence bounds are strongly dependent on the prior distribution, which can lead to dramatically anti-conservative confidence bounds (see Figure 5.3.4).

4.2.3. Sparse Estimation. When the statistical model is not identifiable it may still be useful, for the purposes of visualization or comparative analysis, to quantify a representative set of transcripts. A natural assumption is that the set of transcripts present in solution for a given gene is small. Optimally, we would identify the smallest such subset of transcripts that with a likleihood near the maximum in the unconstrained model, but this is not computationally feasible. Instead, we maximize the augmented objective, $\max_j \left\{\sum_i Y_i \log\left[X\vec{t}\right] - \frac{\lambda}{t_j}\right\}$ subject to $t_j \geq 0$ and $\sum_j t_j = 1$, where λ is a tuning parameter that determines the sparsity

of the resulting solution[42]. Although this optimization problem is not convex, it can be solved by solving N_t convex problems.

We wish to choose the largest λ that guarantees that the sparse solution, $\vec{t}$, lies with the confidence region, Δ_R. That is, we choose λ such that $\sum_i Y_i \log\left[X\vec{t}\right] - \frac{\lambda}{\left\|\vec{t}\right\|_\infty} \geq \sum_i Y_i \log\left[X\hat{t}^\star\right] - \frac{1}{2}\chi_1^2(2\alpha)$, where $\hat{t}^\star$ refers to the maximum likelihood solution, and we use 2α because the confidence bound test is one-sided. Setting $\left\|\vec{t}\right\|_\infty$ to $\max\left\{\min\ \vec{t}\right\}$, the maximum lower confidence bound, $\lambda \leq \frac{1}{2}\max\left\{\min\ \vec{t}\right\}\left[\chi_1^2(2\alpha) - \chi_1^2(\alpha)\right]$. Even though λ is typically very close to 0 in the unidentifiable case, in such cases very small values of lambda can change the solution substantially because a large portion of the parameter space has likelihoods very close to the maximum.

4.2.4. Simulations. We used the simulation script distributed with GRIT to simulate mapped read data for all simulations. The tool works by first sampling a random transcript from the provided frequency distribution, then sampling a random fragment length from the provided fragment length distribution, and finally choosing a fragment uniformily from the chosen trancript with the chosen fragment length until the desired number of samples is achieved. We do not introduce any sequencing or mapping artifacts into the simulated reads. We note that this simulation is consistent with both the GRIT, Cufflinks, and Rsem transcript expression models.

4.2.4.1. *Synthetic Gene Simulations.* For each simulation we sampled from the transcripts uniformly, with a Normal(150, 25) fragment length distribution truncated at ± 2 standard deviations. We ran GRIT, Trinity, and Cufflinks, and used compare_annotations.py with a boundary match of ± 20 basepairs to calculate sensitivity and specificity numbers, which is distributed as part of the GRIT software package. We ran 100 simulations total; 20 simulations with each of 100, 1000, $1e^4$, and $1e^5$ simulated reads.

4.3. Complexity of the Human Transcriptome

Estimating transcript expression from short read RNA-seq data is not in general possible because of the identifiability problem. As seen in 4.2.4, even in simulated data, the current generation of tools fail for simple gene models. A few genes, like DSCAM, are known to produce transcript sets that can not be precisely quantified using short read RNA-seq data. However if most genes produce relatively few transcripts, as is the current dogma, then it is still possible to estimate transcript expression under sparsity assumptions.

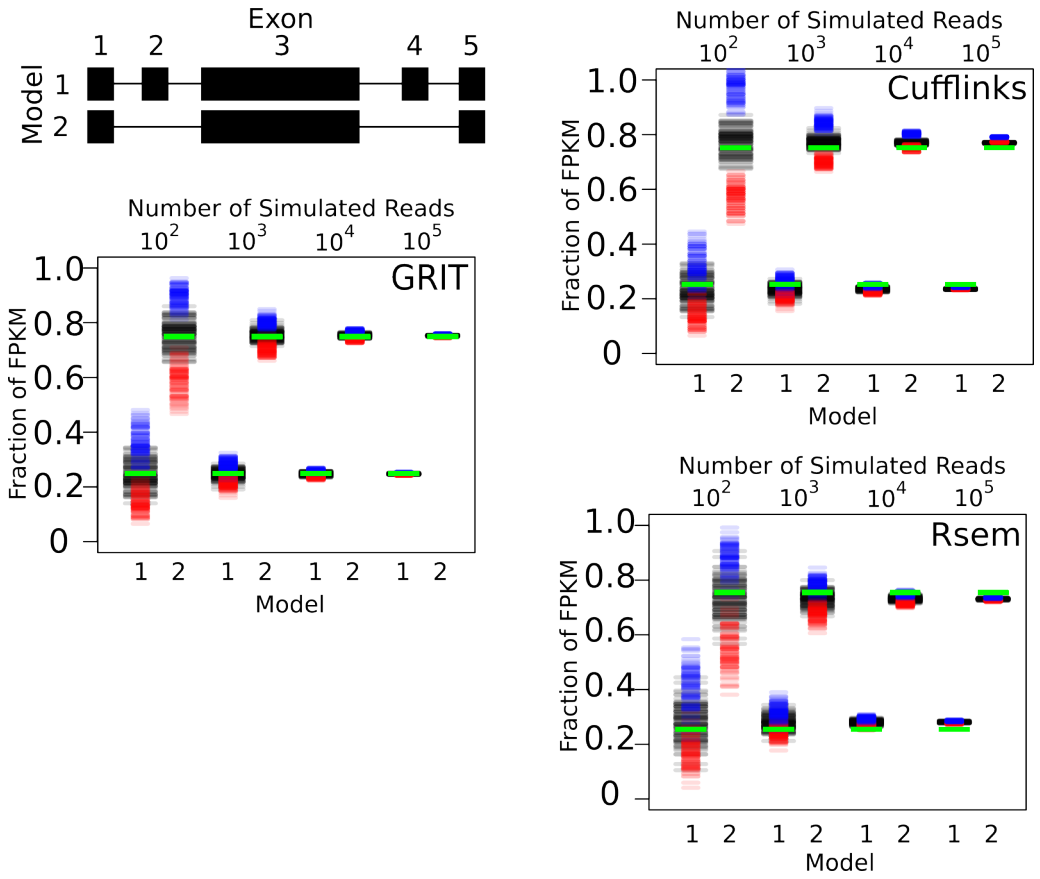

FIGURE 4.2.2. **Identifiable Simulations:** We simulated from models 1 and 2, with frequencies of 0.75 and 0.25 respectively. All methods perform reasonably well, although Rsem and Cufflinks estimates exhibit a slight bias.

Recently long read sequencing technologies have matured to the point where it is possible to directly interrogate the transcriptome. We analyzed unfragmented RNA-seq data that was sequenced on a PacBio sequencer (http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html). Specifically, we identified all pairs of cassette exons in GENCODE version 19 that could be co-expressed, were more than 800 basepairs apart in every GENCODE annotated transcript, and were expressed with an average read coverage of at least 1.

If it were true, as many believe, that the transcriptome were relatively sparse then we would expect to observe only two combinations of a particular pair of cassette exons in the majority of genes. However, of the 28,674 pairs of cassette exons identified we observed all 4 combinations in 62% of transcripts (see Figure 4.3.1). This suggests that alternate splicing events greater than 800 bases apart are regulated independently. It also suggests that precise identification of transcript frequencies is not possible given current technologies.

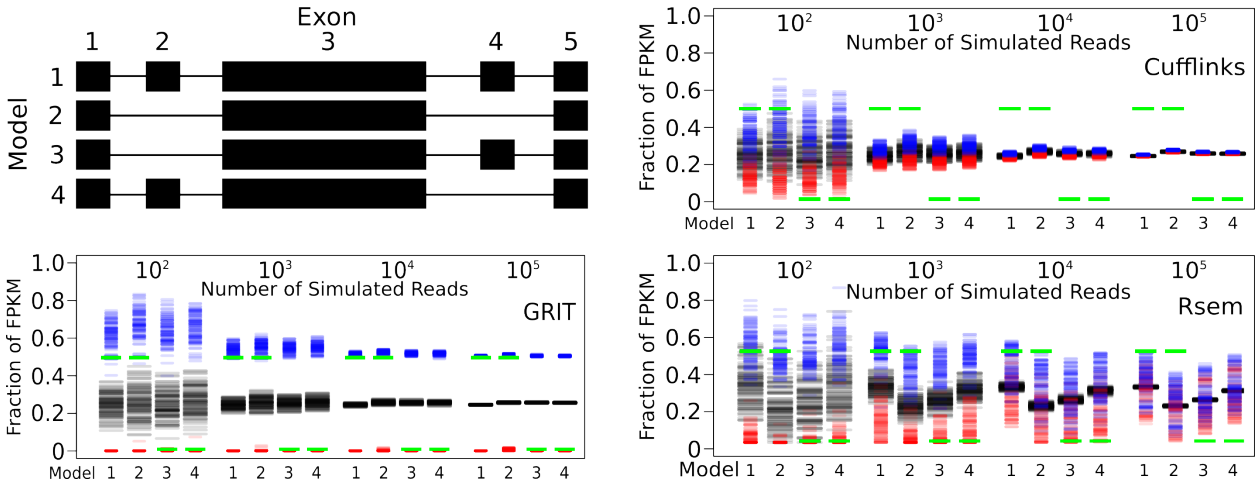

FIGURE 4.2.3. **Unidentifiable Simulations:** We simulated from all four models, with frequencies of 0.49, 0.49, 0.01, and 0.01 for models 1-4 respectively. The green bar is the true freqeuncy. Blue bars identify estimated upper bounds, black bars represent estimated frequencies, and red bars represent estimated lower bounds. Because of the identifiability problem, no methods are able to correctly estimate the transcript frequencies. However, only GRIT is able to properly estimate the confidence bounds.

4.4. Element Expression Estimation

Although it is not possible to precisely identify transcript expression in many studies the quantity of interest is the expression of a gene, or gene element. For example, if one were studying a cancer mutation which produced a receptor variant leading to unrestrained cell proliferation, we may wish to quantify the fraction of transcripts that contained that receptor mutation. Similarly, one may be interested in the sum of the concentrations of all transcripts produced at a particular gene locus. Here, we show that it is possible to precisely identify such quantities even when the individual transcript expression values can not be estimated.

4.4.1. The "select first" paradigm. We follow the lead of popular tools such as Cufflinks and eXpress, modeling an RNA-seq read as being obtained by first selecting a fragment length, then selecting a transcript, and finally selecting a start position within that transcript (see 3). Hence, the probability of obtaining a fragment overlapping any particular exon in the transcript is given by:

$$P[f\ e] \quad = \quad \sum_t P[f \sim t] P[f \sim e | f \sim t] \tag{4.4.1}$$

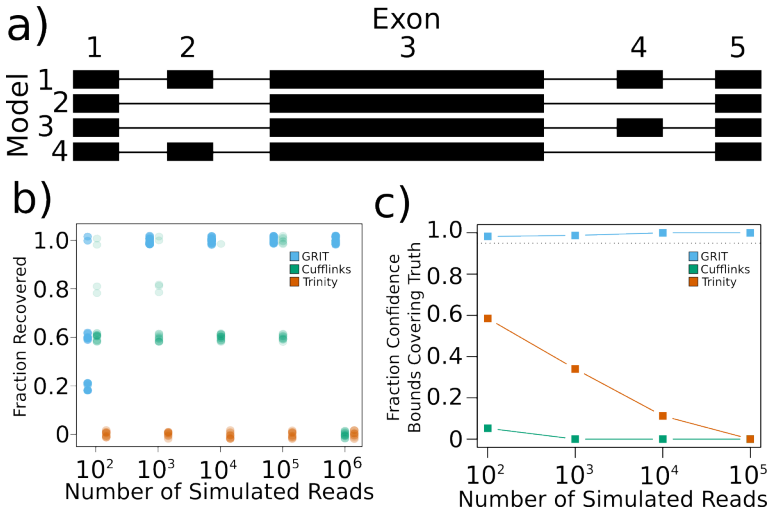

FIGURE 4.2.4. **(a) Simulation Models:** The set of transcript models we simulated from for figure panels **c** and **d**. Because the middle exon is 600 basepairs - longer than the length of the largest fragment - it is impossible to observe exons 2 and 4 in the same read. Thus the staistical model is not identifiable when all four transcript isoforms are present. **(b) Transcript Recovery:** We simulated reads in equal proportions from all four models in panel **b**, and found that only GRIT is able to consistently recover all four models with over a thousand reads. Trinity did not correctly recover any transcript models. Cufflinks recovered 2/20 with 100 reads, 2/20 with 1000 reads, 1/20 with 10k reads, and 6/20 with 10 thousand reads. However, because of the shortest path assumption, each time it built all four models it created an artificial TSS or TTS between 20 and 50 basepairs from the true TSS or TTS. When we restricted the transcripts to be equivalent only when the gene boundaries are within 10 basepairs of the truth, then Cufflinks did not correctly identify more than two models correctly. **(c) Confidence Bound Accuracy:** We simulated reads from all for models in panel **b**, with frequencies of 0.49, 0.49, 0.01, and 0.01 for models 1-4 respectively. For each tool, we plotted the fraction of times that the estimated confidence bounds contained the truth. The dashed black line is at 0.95, the expected fraction of times that the confidence bounds should contain the truth. GRIT's confidence bounds are slightly conservative, covering the truth an average of 99% of the time. Because of the identifiability problem, Cufflinks and Rsem confidence bounds are extremely anti-conservative, never covering the truth for n=10000. This is a summary of the data plotted in panel **f**. Note that, because over 30% of genes have both alternate TSS's and alternate TES's, Cufflinks and Rsem have the potential to produce anti-conservative confidence bounds for a large fraction of annotated gene loci.

If we further assume that we can uniquely identify fragments from sequenced reads and that reads are uniformly distributed across transcripts, which is pedagogically useful, we have:

$$\begin{aligned} P[r \sim e] &= \sum_t \frac{P[r \sim t] L(e)}{L(t)} \\ &= \sum_t \frac{N_t L(t)}{N_W L(w)} \frac{L(e)}{L(t)} \\ &= \sum_t \frac{N_t L(e)}{N_W L(w)} \\ &= \frac{L(e)}{N_W L(w)} \sum_t N_t \end{aligned} \tag{4.4.2}$$

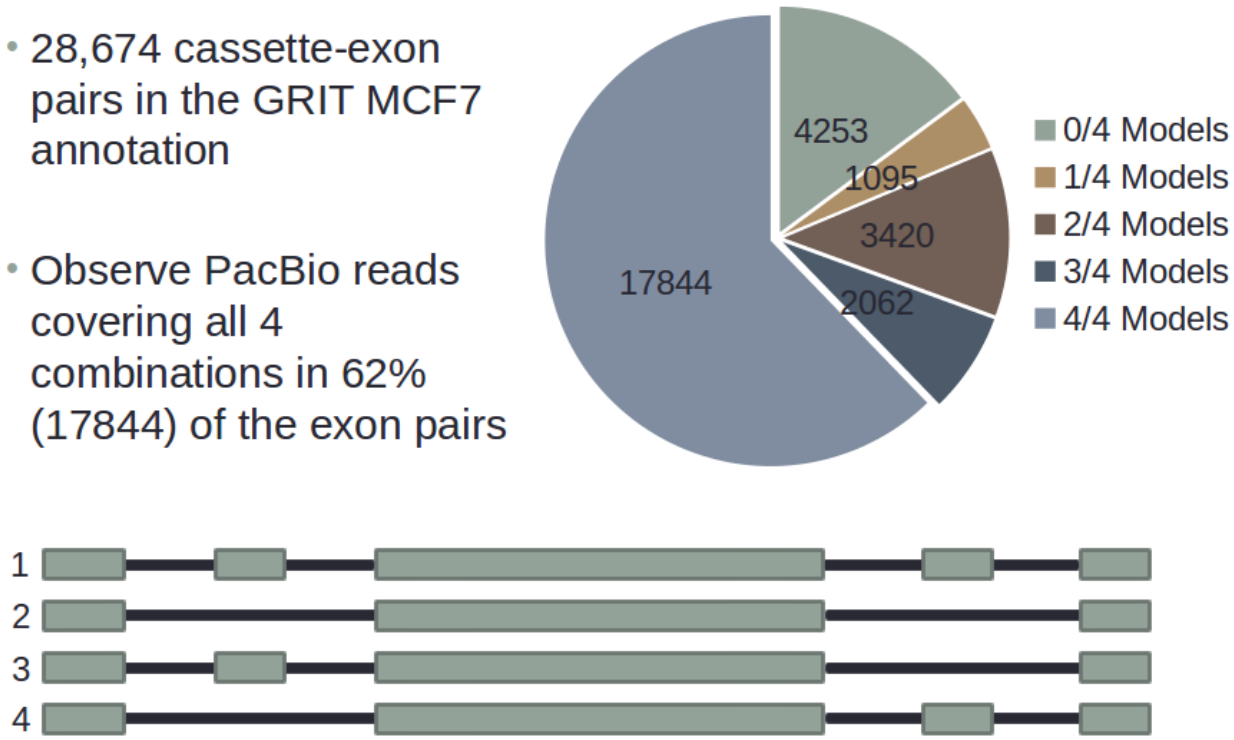

FIGURE 4.3.1. We identified all pairs of cassette exons in GENCODE version 19 that: 1) could be co-expressed 2) are more than 800 basepairs apart in every GENCODE annotated transcript 3) are expressed with an average base coverage of 1. Of the 28,674 pairs identified, we observed all 4 combinations in the majority of exon pairs.

where the sum in w ranges over all transcript isoforms in the library, $L(\cdot)$ is the number of distinct fragments that can overlap an element, and N_t is the count of transcripts of the isoform type in the library. Note that the ratio $L(t)/L(e)$ is equal to the ratio of the lengths (in base pairs) for $L(f) \lll L(e)$.

Models similar to equation 4.4.1 underpin Cufflinks ([53]), RSEM ([35]), eXpress ([45]) and many other widely used software packages. These packages differ in the details of how they model the probabilities in 4.4.1, but each reduce down to a form similar to 4.4.2 where it is necessary to know the length of every transcript isoform in the RNA library in order to compute the expression level of any one. In practice, we rarely have reliable structural information about even a small subset of transcripts, particularly about their lengths, which requires precise information about transcript start and end sites. However, note critically that 4.4.2 does not depend on the length, $L(t)$, of the transcript in question except through the normalizing sum in the denominator. We will exploit this shortly.

The count of exons in the library, N_e, is equal to the sum of the count of transcripts containing that exon, $\sum_t N_t \mathbb{I}[t \in e]$. Hence

$$\begin{aligned} P[r \sim e] &= \frac{L(e)}{N_W L(w)} \sum_t N_t \\ &= \frac{L(e)}{N_W L(w)} N_e \end{aligned}$$

Note that N_e does not specifically depend on the transcripts which contain it, but rather on the marginal distribution of transcripts lengths in the RNA library. Thus N_e is only weakly correlated with the length of the transcripts in which the exon occurs.

In addition note that, since we only care about the sum, if the total amount of RNA in the library, $\sum_t N_t$, is known and the length and count of even a single transcript isoform is known, then the normalizing factor $\sum N_w L(w)$ can be estimated by $\frac{N_t L(t) M}{P[r\ t]}$. Spike-ins provide a direct method to estimate this quantity.

The generative model is conceptually simple, but can lead one to overestimate the necessity of complete inform about the set of transcript isoforms for the purposes of quantification.

4.4.2. The "fragment first" paradigm. Here we model an RNA-seq read as obtained from a pool of fragmented RNAs; transcripts are never expressly selected. It is clear that the probability of selecting a read from a given transcript will still depend on its length, and that selecting a read from a particular isoform (a set of identical transcripts) will depend on the abundance of the set compared to other isoforms in the library. This variance relation is not apparent in simple pedagogical examples, where the entire library consists of only a few transcript isoforms.

This has the consequence that the relative expression of exons can be estimated directly from reads overlapping the exon, without appeal to full length transcript models. This observation also helps to explain why "wiggle" tracks displaying local read coverage are useful for exploratory analysis: they encode the majority of the information available from the RNA-seq experiment (at least in real world, complex library scenarios). Similarly, this emphasizes the utility of count-based methods for differential expression analysis, and may help to explain some of the generally very good concordance between transcript level differential expression analysis and those based on counts. We note that our findings are consistent with the model behind underlying Cufflinks, and with additional assumptions, DE-seq, edgeR, Voom, MISO, DEx-seq, and most other statistical models in use.

4.5. Discussion

We have presented a comprehensive model of transcript element expression which makes minimal assumptions about transcript structure, precisely estimating transcript expression when necessary and producing conservative conservative bounds when not. In addition we show that, even when it is not possible to identify transcript frequencies, it is possible to estimate transcript element and gene expression. We have developed a tool, GRIT, which implements these estimation procedures. GRIT can be downloaded at http://grit-bio.org/.

CHAPTER 5

Transcript Discovery

5.1. Introduction

The practice of sequencing short fragments of cDNAs on a next or third generation sequencing platform is known as RNA-seq. This assay yields quantitative information about gene expression, novel alternative splicing events, RNA editing, poly-adenylation sites, and other phenomena [57, 21, 40]. Since its inception, the prospect of utilizing RNA-seq data to "assemble" new gene and transcripts models has motivated the development of algorithms and software [53, 19, 22, 50, 46]. De Novo assembly methods, like Trinity[19], Oases[50], and Trans-Abyss[46] align reads to construct transcript sequences, which can then be mapped to a reference genome. Genome guided approaches, like Cufflinks[53] and Scripture[22], use reads that have previously been aligned to a reference genome to identify transcript models.

The impact of RNA-seq data on genome annotation has been most substantial for new or non-model organism. For instance, RNA-seq data and Cufflinks were used to produce a de novo annotation for the sea urchin, Strongylocentrotus Purpuratu, but incorporated a stringent filtering system that removed novel transcript models that lacked ORFs longer than 500aa or that didn't encode a known protein [54]. In organisms with more established reference transcriptomes, the impact of RNA-seq data largely been via manual incorporation of elements discovered from RNAseq data. GENCODE, the annotation group within the ENCODE Consortium, has only used PCR-validated RNA-seq splice junctions along with 5' end information from Cap Analysis of Gene Expression (CAGE) data to conduct manual annotation. FlyBase [17], the primary annotation effort for Drosophila, has used RNA-seq data to manually modify gene models inferred from full length cDNAs and RNA-seq discovered junctions. Ensembl used RNA-seq data to modify and extend the annotation of the zebrafish genome [9], but the effort discovered few novel full-length transcript models. RNA-seq data was combined with trans-spliced leader sequences and polyadenylation tracts to improve the quality of the Wormbase annotation of Caenorhabditis elegans, but a lack of high quality annotation tools led the authors to develop their own analysis techniques and ultimately they focused their analysis on transcript elements, rather than full length transcripts.

It is not surprising that full length transcript annotation has primarily remained in the domain of manual annotation and full-insert cDNA sequencing, because RNA-seq reads are too short to cover full transcripts, typically providing information only about three or four exons at a time [31]. This means that it is not always possible to positively identify alternate transcript isoforms, even as the read depth approaches infinity. Furthermore, biases in the RNA-seq assay make positive identification of novel transcript boundaries difficult[5, 23, 57, 44]. Other annotation tools attempt to circumvent these problem by placing additional restrictions on the space of discoverable transcripts. For instance, Cufflinks only permits the minimal set of transcripts needed to explain the splice junctions, over-simplifying complex loci like Dscam. Trinity always extends transcript contigs to the last base, disallowing nested promoters and nested poly(A) sites. As we show, these restrictions can produce annotation sets that are in direct contradiction to observed data from complementary assays.

We introduce a new method, Generalized RNA Integration Tool (GRIT), which we show performs better than competing tools by utilizing a novel statistical technique combined with the integration of gene boundary data. Our approach allows for the construction of any transcript models that can be built by Cufflinks, Trinity, Scripture, Oases and Trans-Abyss, although our requirement that every transcript model be supported by experimental evidence can make it more restrictive in practice. For the purposes of benchmarking, we have utilized a subset of the modENCODE dataset (1.67B bp) to compare the performance of GRIT to the most widely used transcript-level RNA-seq analysis tools. GRIT has also been applied to the full set of modENCODE RNA data (over 1 Terrabase of sequence data) to generate a data driven annotation of the fruit fly transcriptome. These gene and transcript models use CAGE, RACE, EST, cDNA, 454, stranded paired-end RNA-seq, and poly(A)+seq, to provide an uprecedentedly detailed look at the biology of eukaryotic genomes. The full length transcript models that we derive reveal, for instance, that over 20% of protein coding genes encode multiple localization signals, and alternative polyadenylation is more common than alternative splicing in neuronal tissue. These and related insights reported here and in Brown et al.[4] were not obtainable with other analysis tools, and underscore the importance of integrating multiple assay types when interpreting Next Generation RNA sequencing data.

5.2. GRIT: A tool for integrative analysis of RNA data

GRIT uses reads aligned to a reference genome to build transcript models. We make few assumptions about the structure of a transcript, but do require that every element (e.g. promoters or splice junctions) is supported experimentally. A "transcript" is a set of genomic regions that begin at a transcription start site (TSS), extend through one or more exons connected by splice junctions, and ends with a transcription end site (TES).

We define four distinct element types: TSS exons, TES exons, internal exons, and single exon transcripts. TSS exons begin with an experimentally detected promoter (e.g. via the CAGE or RACE assays or 5' EST sequencing [29]), and end with a splice donor site. Similarly, TES exons begin with a splice acceptor site and end with an observed TES (e.g. a poly(A) site). Internal exons begin and end with a verified splice site, and single exon transcripts begin with TSS and end with a TES. Our transcript models can use both canonical and non-canonical splice sites. The set of candidate transcripts includes both single exon transcripts and transcripts that begin with a TSS exon, contain splice-junction connected exons, and end with a TES exon (Fig 1b - Exon Graph).

The GRIT annotation pipeline consists of four parts described below: gene region identification, element discovery, transcript construction, and transcript expression estimation.

5.2.1. Identifying Gene Regions. Segmenting the genome into gene regions involves three distinct steps: indentifying exonic regions, identifying intronic regions, and merging exonic and intronic regions into gene regions.

To build a set of exon regions, we scan the genome for 100 basepair regions without any RNAseq, CAGE, or poly(A)+seq reads. These empty regions form boundaries between the different exonic regions.

To identify introns, we collect reads that map in a non-contiguous fashion to the reference genome, typically known as junction reads. To avoid junction reads that may be experimental or mapping artifacts, we filter the set of identified junctions using the filtering method described in [21], which requires that junctions have an entropy defined as:

$$\begin{aligned} p_i &= \text{reads at offset } i/\text{total reads to junction window} \\ Entropy &= -\sum_i p_i \log_2(p_i) \end{aligned}$$

We require that junctions have an entropy score of at least two in one biological sample. In addition, although the RNA-seq assays we analyzed are highly stranded, there is some low-level unstranded background. Hence, to remove incorrectly stranded reads, we remove junctions on the strand opposite of canonical acceptor / donor sequences if their frequency is less than 5% of the junction frequency on the canonical strand. The junction reads that pass this filter are then aggregated into a set of discovered introns.

Finally, we merge exon regions that share a discovered intron, forming our gene regions. Note that, although 100 basepairs is too large to properly separate many gene pairs, in practice it provides a good first approximation. During the element discovery stage we use the identified

CAGE and poly(A)+seq peaks in combination with the read coverage to further segment when necessary.

5.2.2. Element Discovery. Element discovery proceeds independently in each gene region. We split the gene region into non-overlapping segments with attached labels that describe the segment boundary (see 5.2.1a). For instance, a segment where the left boundary is a splice donor and right boundary is a splice acceptor is a canonical intron; a segment where the left boundary is a splice acceptor and right boundary is a splice donor is a canonical exon. There are four boundary labels: splice acceptor, splice donor, TSS, and TES. Splice donors and acceptors are identified directly from junction reads; TSS and TES are identified from CAGE and poly(A)+seq data using the method described in Chapter II.

All possible pairwise combinations of the four segment boundary labels produce 16 possible combinations, which we group into seven segment labels: TSS segments, canonical introns, canonical exons, exon extensions, TES segments, single exon transcripts, and intergenic segments (see 5.2.1a - Labeled Segments). TSS segments are any segments with a left TSS label; similarly, TES segments are segments where the right boundary has a TES label. Canonical introns have a left splice donor label, and a right splice acceptor label. Canonical exons have a left splice acceptor label, and a right splice donor label. Exon extensions either have two splice donor labels, or two splice acceptor labels. Single exon transcripts have a left TSS label, and a right TES label. Labeled segments with low read coverage segments are now removed.

Within a gene region, a low coverage region is defined as a segment where the average read coverage is lower than a global threshold (1e-2) with high probabilility; or, the ratio of a segments average read coverage to the highest read coverage segment in the same gene region is less than 1% with high probability.

Regions that begin with a left TES label and end with a right TSS label are intergenic segments. If intergenic segments are discovered and the average base coverage is sufficiently low, then the gene region is split and the element discovery process is re-started in both halves.

The set of candidate exons is all combinations of adjacent segments that start with TSS or splice acceptor, and end with a TES or splice donor. Regions that begin with a TSS label and end with a donor junction are TSS exons; regions that begin with a acceptor junction and end with TES label are TES exons; regions that begin with a donor junction and end with an acceptor junction are internal exons; regions that start with a TSS label and end with a TES label are single exon transcripts (see 5.2.1a - Exons).

5.2.3. Transcript Discovery. For the purposes of candidate transcript construction, we model a gene as a directed graph in which each exon is a node, and splice junctions are edges (see 5.2.1b - Exon Graph). Then the set of candidate transcripts is all possible paths through this graph that begin with a TSS exon and end with a TES exon (see 5.2.1b - Example Path). This differs from other methods, e.g. Cufflinks, in that we consider all possible paths subject to this restriction, rather than some minimal set of covering paths.

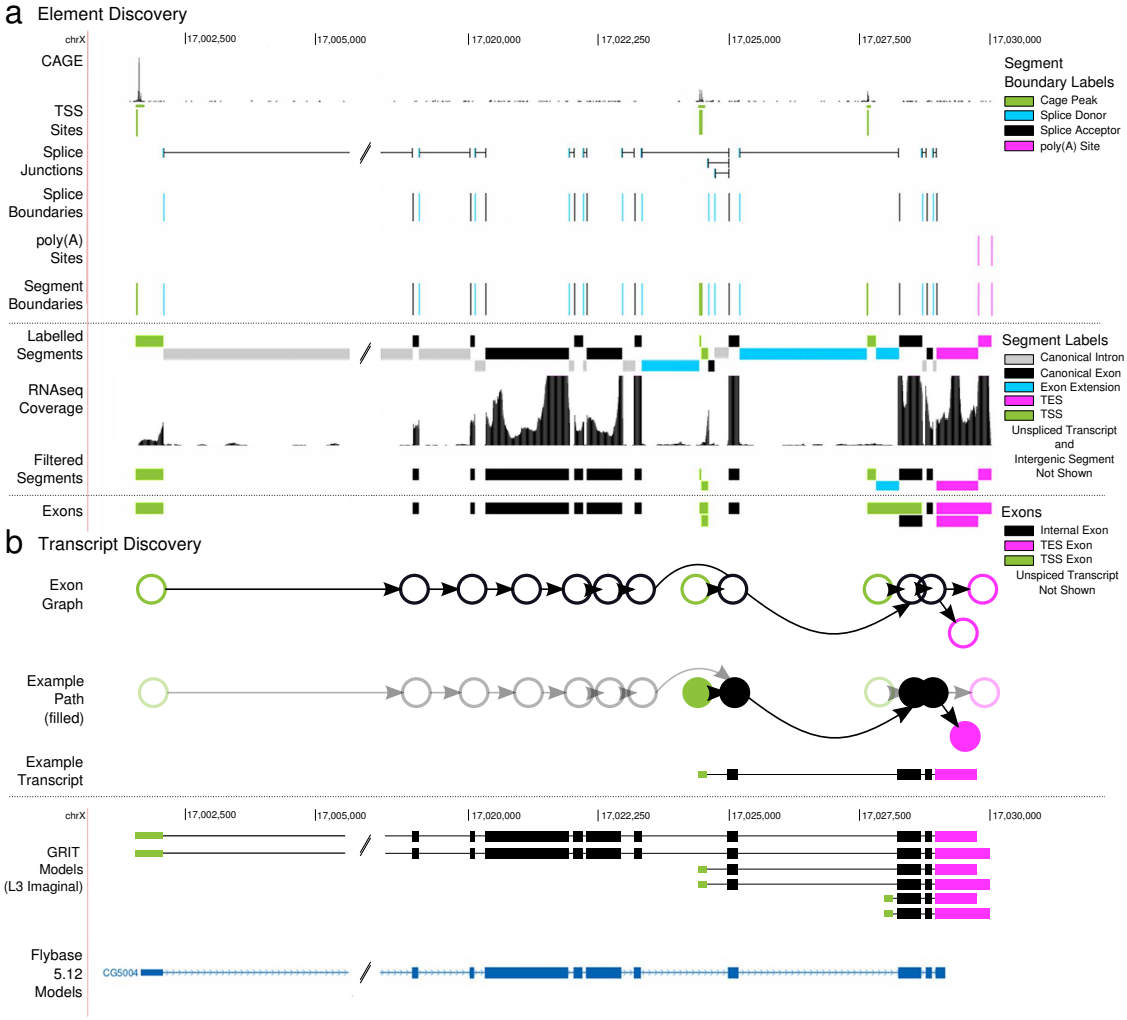

FIGURE 5.2.1. **Element Discovery Overview** - See Section 2.1.1-2.1.4 for a more detailed description **(a) Exon Discovery:** For each gene segment we identify CAGE peaks; segment the gene region using the CAGE peaks, splice boundaries and poly(A) sites; label the segments based upon their boundaries; filter intron segments with low RNA-seq coverage; and build labeled exons from adjacent segments. **(b) Transcript Discovery:** For each gene, we construct a graph where each node is an exon discovered in (b), and each edge is a junction. Then, each candidate transcript is identified with a single path through this directed graph that begins with TSS node, and ends with a TES node.

5.2.4. Sensitivity To Tuning Parameters. GRIT has two main tuning parameters: one that governs the threshholding of segments with low read coverage, and one that governs the retention of canonical introns.

Changes to the minimal exon read coverage tuning parameter affects the results very little over reasonable ranges. For instance, in the data set we analyzed for the purposes in this manuscript, changing this parameter from 0.01 BPKM to 1 BPKM reduces the sensitivity by less than 1%, and increases the specificity by less than 1%. This is consistent with our observation that the limiting factor for transcript construction is junctions reads, rather than read coverage within a gene body.

The other important tuning parameter is the canonical intron retention threshold and, unfortunately, the optimal value is a function of the assay type. For instance, we have applied GRIT to total RNASeq (data not shown) and find that a threshold of 80% percent is necessary to prevent the routine inclusion of unprocessed elements. However, in the poly(A)+ data that we analyzed for this study, a threshold of 5% was sufficient to exclude the vast majority of unprocessed transcripts. We currently err on the side of conservatism, setting this to 80% by default. This setting has the potential to miss retained introns in poly(A)+ RNASeq, but seems to provide good results over a wide variety of organisms and protocols.

5.3. Comparison to Competing Tools

Current transcript discovery tools make assumptions about the structure of the underlying transcripts, usually restricting them to some identifiable subset. For instance, Cufflinks assumes that the set of possible transcripts is the minimal set of covering paths in the graphical model described in Section 2.1.3. Trinity requires that transcript models extend to the furthest base of an assemblable contig, which disallows transcript models with nested transcription start and termination sites. The GRIT model allows for both of these, but requires gene boundary information.

5.3.1. GRIT Discovers More FlyBase Transcripts with Higher Precision. We benchmarked GRIT against Cufflinks, Scripture, and Trinity+Rsem. We used stranded RNA-seq, CAGE, and poly(A)+seq data produced from dissected heads of 20 day adult flys.

We analyzed the recall and precision of the transcriptomes generated by each tool by comparing them to the expressed 7079 FlyBase 5.45[39] genes (13141 transcripts) (Supp 1.5.1). Transcripts were considered equivalent when they had the same internal splicing structure and gene boundaries within 50 basepairs of each other. Under this measure, GRIT recovers 44.2% of transcripts with 17.8% specificity versus 13.4%/8.8% sensitivity/specificity

for Cufflinks, 8.6%/3.0% sensitivity/specificity for Trinity+Rsem, and 0.9%/1.4% sensitivity/specificity for Scripture (Figure 5.3.1). When we filter predicted transcripts with an expression score less than 1e-6 estimated fragments per kilobase per million reads (FPKMs) at a marginal 99% confidence level, then GRIT recovers 39.8% of FlyBase transcripts with 41.3% specificity. The Cufflinks, Trinity, and Scripture numbers are essentially unchanged.

This substantial rise in specificity is largely due to eliminating complex genes. The GRIT annotation is heavily penalized in complex loci, e.g. Dscam1 or Mhc, because FlyBase includes new transcript models when they contribute a novel exon, intron, or gene boundary (Flybase 5.45 gene notes). GRIT's superior performance is not purely a result of its increased ability to precisely predict transcript boundaries; when we relax the transcript boundary match distance to 200 basepairs, GRIT still out-performs competing methods (Figure 5.3.1).

We studied the consistency of transcript expression scores estimated by calculating the correlation between estimated FPKMs and both CAGE and poly(A)+seq tag counts. GRIT annotated transcripts are able to achieve Spearman rank correlations between 0.71 and 0.80 across replicates, while Cufflinks, Trinity, and Scripture correlations are all below 0.5 (Figure 5.3.1).

To study the specificity of the annotated transcript boundaries, we analyzed the motif enrichment of the two most spatially localized core promoter motifs, TATA[37] and Inr[6], in regions within 50 bases of annotated TSSs (Figure 5.3.1).

To identify motif enrichment in the genome sequence surrounding annotated TSSs, for each tool, we first identified the unique set of transcript start sites. Then, for each TSS, we scanned the genome sequence taken from BDGP5 genome for the the TATA motif (TATAAA) and the Inr Motif ([CT][CT]A[ACGT][AT][CT][CT]). A base position was considered a hit if the motif match was exact. Finally, we summed the number of hits at each position, and then divided by the total number of sequences to produce enrichment numbers. To identify significantly enriched regions, we used a non-parametric approach, performing the above analysis 10000 times using sequence chosen randomly from transcribed regions throughout the genome. A particular position was considered enriched if it's value was greater than 9999 of the bootstrapped samples. 9999 was chosen so that the Type I error rate under the NULL is expected to be 1%, after accounting for multiple testing.

The genome sequence surrounding GRIT and Scripture identified TSSs are significantly enriched for the TATA motif 24-32 and 30-35 bases upstream of the TSS, respectively. These correspond to 3.2% and 1.1% of distinct annotated TSSs. Regions identified by Cufflinks and Trinity are not signficiantly enriched for the TATA motif at any positions. Similarly, GRIT identified regions are significantly enriched for the Inr motif enrichment at ± 1 bases

of the TSS, which corresponds to 12% of annotated TSSs. Neither Cufflinks, Trinity, nor Scripture identified regions are significantly enriched at any bases for the Inr motif.

We also analyzed the regions within 50 basepairs of FlyBase 5.45 annotated TSSs, and found TATA enrichment at 27-34 bases corresponding to 2.9% of distinct TSSs, and Inr enrichment 2-3 bases upstream of annotated TSSs, corresponding to 1.5% of distinct annotated TSSs. Although both GRIT and Flybase TSS regions show similar TATA enrichment, GRIT more precisely identified the 26-31 basepair upstream positioning[6]. The GRIT enrichment results are consistent with previous studies [29], which report TATA and Inr motifs in 2.1% and 13.8% of peaked promoters identitifed by RACE[30].

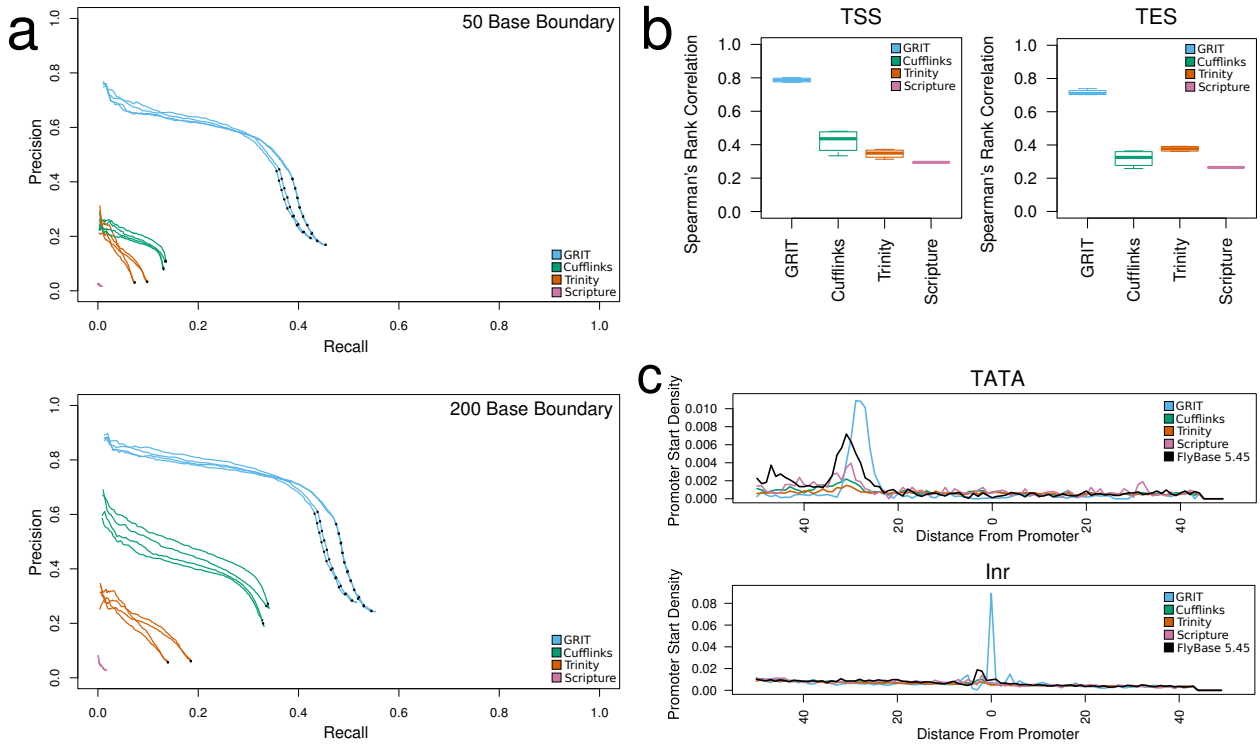

FIGURE 5.3.1. **Comparison with Existing Tools**: **(a) Sensitivity and Specificity Analysis:** We compared the set of transcript isoforms discovered by GRIT, Cufflinks, Scripture and Trinity to the FlyBase annotation. A transcript was identified as a match if the internal structure was the same, and the distal boundaries were, variously, within 50 and 200 of one-another. **(b) FPKM versus CAGE and poly(A)+seq Counts:** For each sample, we calculated the Spearman Rank Correlation between estimated transcript FPKMs and raw CAGE and poly(A)+seq read counts within 50 bases of each annotated promoter/poly(A) site. **(c) Motif Analysis:** For each sample, we considered the sequence within 50 bases of annotated promoters. A position was considered a TATA motif hit if it matched the sequence "T-A-T-A-A", and an Inr motif match if it matched the sequence "C/T-C/T-A-N-A/T-C/T-C/T". The plots are aligned with respect to the first base in the annotated promoter, and plot the fraction of promoters that contain a motif match at each position, averaged over replicates.

5.3.2. Alternate transcript boundaries are common and differentially regulated. Alternate promoters have long been known to serve a regulatory role. The sequence of both 5' UTRs introns within 5' UTRs have the potential to alter translational efficiency and subcellular localization of the mRNA. Alternative N-terminal protein sequence is known to control the localization of many proteins.

Genes encoding alternative N-[terminal domains, either by alternative promoter usage or splicing, include well-studied examples such as the prothoracicotropic hormone (Ptth) gene critical for metamorphosis in insects[33, 43]. Ptth encodes three neural-secreted hormone protein isoforms: the canonical form containing a signal peptide sequence for exportation from the cell; a second isoform with a 25 amino acid N-terminal extension containing a mitochondrial targeting peptide; and a third form which is shorter than the canonical isoform by nine amino acids (Figure 5.3.3). The third short isoform is predicted to be localized to the cytoplasm or nucleus. Ptth encodes for multiple localization signals, which appear to be a general phenomenon: we find that 33% of alternative start codons encode known alternative protein localization signal (compared to 4.6% of internal cassette exons $p < 1e\text{-}100$ by t-test, and 14% of alternative C-terminal coding sequence, $p < 1e\text{-}40$ by t-test). Since the majority of known localization signals are N-terminal, the enrichment relative to other alternative exons makes a useful negative control. The remarkable observation is that nearly 20% of all protein coding genes encode multiple localization signals.

We also find substantial complexity at the 3' ends of transcripts. For 77 genes alternative polyadenylation sites alter coding capacity by superceding the stop codon, and has been shown to have the ability to either change the translated reading frame or generate non-coding transcript variants [13].

5.3.3. Correctly Identifying Transcript Boundaries Requires Additional Data. Biases in RNA-seq read coverage have been widely reported [44, 5, 23, 57] and, although several methods have been developed to attempt to remove such bias [36, 61], the methods are typically aimed at correcting transcript expression levels rather than correcting read coverage estimates. As such, local, random changes in read coverage make it difficult to determine whether a particular site is a transcript initiation/termination site, or a random fluctuation in read coverage. To compound this problem, even when we restrict our attention to polyadenylated transcripts so that we can use poly(A) spanning reads to identify transcript ends, sequencing bias make the poly(A) spanning reads much more rare than other read types [10]. For instance, in the modENCODE poly(A)+ data sets, poly(A) spanning reads were roughly 100 times less likely than they would have been were the read distribution over transcripts uniform.

To demonstrate the confounding effect of read coverage bias on transcript boundary identification, we use the CAGE and poly(A) data to determine the extent to which one could identify TSS and TES sites purely from RNA-seq data. For each gene in Flybase 5.45 (FB5.45) with a BPKM greater than 10, we found the 10 basepair window with the highest amount of CAGE signal, and recorded the ratio of the net base coverage 50 basepairs upstream of the site to 50 basepairs donwstream of the site. We calculated the same statistic for the furthest poly(A) site in each gene. These two sets gave us our positive control set. Next, for each gene, we uniformily sampled 10 random locations from within annotated transcription regions, and calculated the signal ratio to build the negative control set. Finally, we estimated the posterior probability of a site being a gene boundary by direct application of Bayes theorem, where the marginal probability of a promoter and poly(A) site were taken from the GRIT identified CAGE and poly(A) regions.

On average, the signal enrichment ratios were 19.7 and 9.7 for TSS and TES's respectively, versus 1 for the negative control set. Using the known frequency of promoters in the genome as an estimate of the probability of a promoter and the estimated enrichment ratios, the maximum posterior probability that a given position is a promoter is 67.9%, and occurs when the upstream to downstream signal ratio is 85.1 Similarly, for poly(A) sites, the maximum posterior probability is 35.5% and occurs when the downstream to upstream ratio is 83.2 (see Figure 5.3.2). Thus, even under ideal conditions, RNA-seq coverage alone is likely insufficient to accurately identify transcript boundaries.

The lack of enrichment for core promoter motifs surrounding TSSs annotated by Cufflinks, Trinity, and Scripture demonstrates the practical challenges in identifying transcript boundaries from RNAseq data alone (see Figure 5.3.1).

5.3.4. Current tools under-estimate splicing diversity. Even if the other tools, e.g. Cufflinks, could be modified to account for transcription start and end site data, they still would not permit the transcriptional complexity that we observe. We found 47 genes with the capacity to encode more than 1,000 transcript isoforms[4], and 27% of these are only present in samples enriched for neuronal tissue. Together, these 13 genes account for nearly 13.5% of the unique transcript isoforms that can be expressed. In Ad20dHeads, 59.6% of genes expressed encode multiple transcript isoforms (Fig 3b). Of these, 29.8% exhibit multiple promoters, 48.1% multiple poly(A) events, and 40.1% exhibit alternate splicing (Fig 3c).

5.3.4.1. *DSCAM Simulation.* Dscam1 is an example of a well studied gene that has the potential to encode 38,016 distinct proteins [7](see Figure 5.3.3), and specific homophilic binding has been observed for over 3000 isoform pairs [59]. DSCAM1 is known to play a

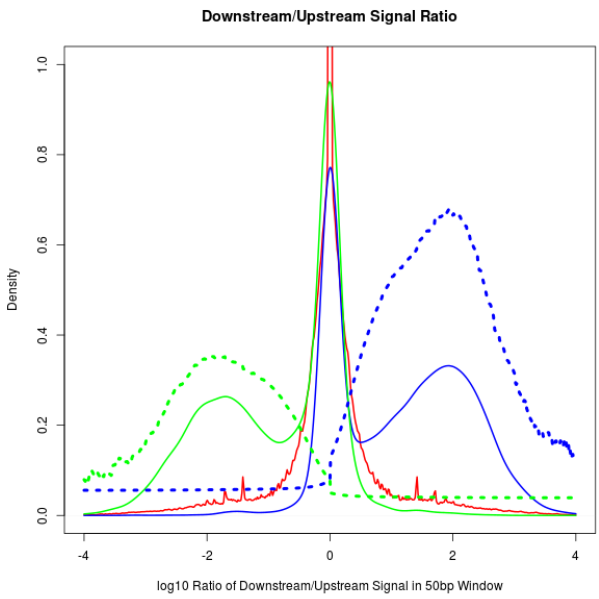

FIGURE 5.3.2. **Identifying Gene Boundaries Solely From RNA-seq:** The dark red line indicates the marginal distribution of RNA-seq signal across exonic regions. The dark blue and dark green lines indicate the distribution of RNA-seq signal ratios over CAGE peaks and poly(A) sites, respectively. The dashed blue and green lines indicate the posterior probability that a location is a TSS or TES, based solely upon its RNA-seq signal ratio. For instance, the dashed blue line peaking at 0.65 indicates that it is impossible to identify a CAGE site from RNA-seq signal ratio alone with greater than 65% certainty.

crucial role in axonal tract formation in the developing fly nervous system, and is expressed in neuronal tissue throughout the lifecycle. We observed the highest levels of expression in the central nervous system of 2 day old white prepupae (WPP CNS), where we are able to identify a 3' extension and two novel cassette exons, allowing Dscam1 to produce as many as 228,096 distinct transcripts. In the data collected from Ad20dHeads, GRIT identifies 720 isoforms with perfect precision, whereas Cufflinks and Trinity were unable to identify a single full length transcript.

We used simulated data to study the ability of GRIT, Cufflinks, and Trinity to recapitulate the complete set of transcript isoforms. We used the set of DSCAM exons from FlyBase 5.45 to enumerate all possible 38016 DSCAM transcript models, and a Normal(300,25) fragment length distribution truncated at ± 2 standard deviations. When fed 10,000 RNASeq reads simulated uniformly from the canonical 38016 isoforms, GRIT was able to recover every transcript isoform with perfect precision in 19 of 20 simulations (see Figure 5.3.4). Trinity was never able to build a full length transcript (see Figure 5.3.4) and Cufflinks was only able to recover a single transcript in 1 out of 100 simulations, demonstrating their inability

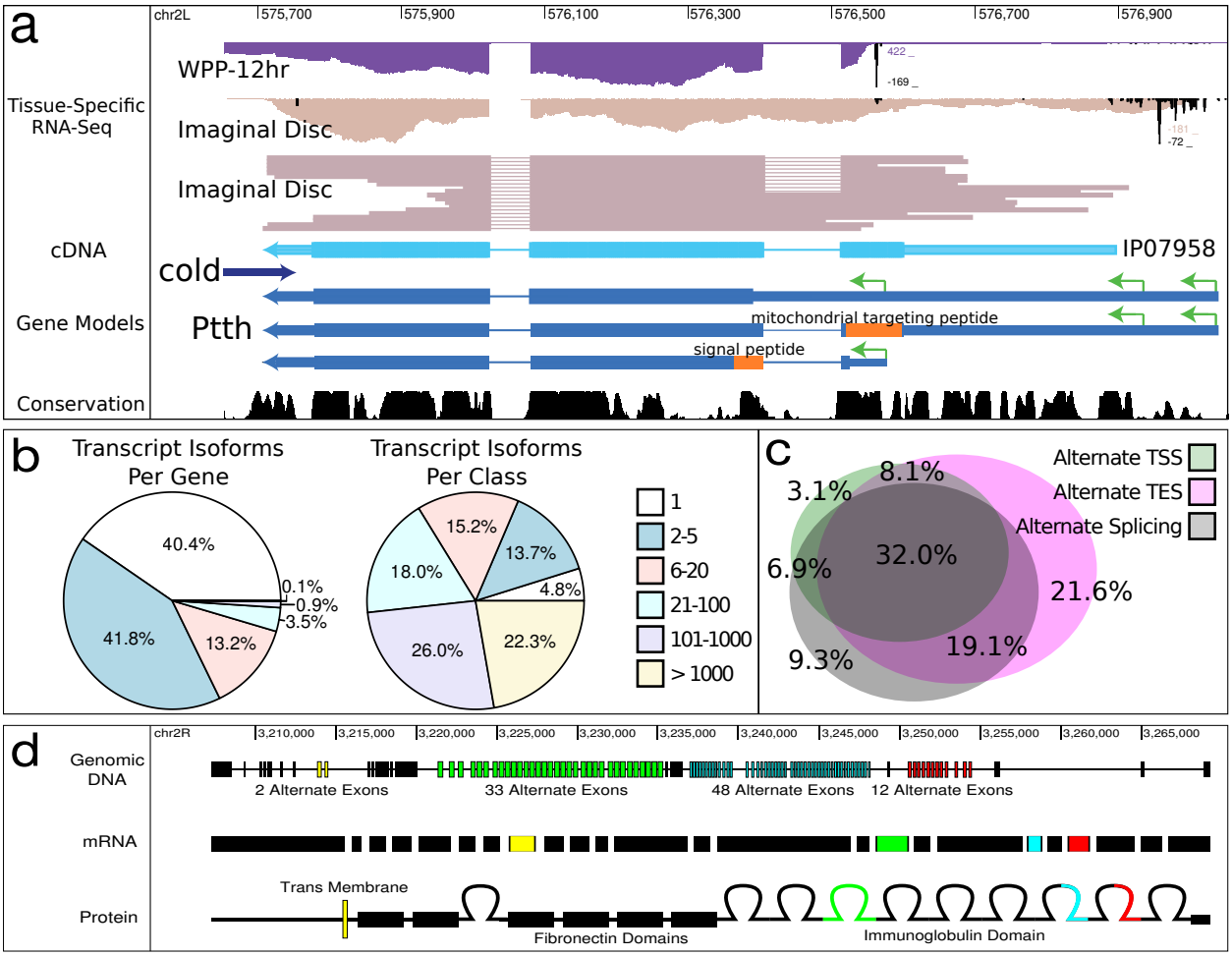

FIGURE 5.3.3. **(a) Ptth:** The Ptth gene encodes isoforms with multiple proteins due to alternative N-terminal splicing as well as promoter usage. The sample labeled "Imaginal Disc" corresponds to mass isolated tissues enriched more than 50% for imaginal discs. **(b) Gene Complexity:** Although most genes have less than five isoforms, nearly half of transcript isoforms originate in genes that encode 100 or more distinct transcripts. **(c) Sources of Gene Complexity:** The Venn digram only represents the 59.6% of genes that encode multiple transcript isoforms. **(d) *Dscam*:** *Dscam* is the most complex gene in *Drosophila*, with most of its complexity coming from the combanatorial inclusion of four sets of 2, 33, 48, and 12 alternate exons. The two cassette exons in yellow moldify the trans membrane domain; the other locus affect the paired binding affinity by modifying the immunoglobulin domains. The figure is essentially as appears in [49].

to account for genes of such complexity (see Figure 5.3.4). Simulating from the 228,096 isoforms identified in WPP CNS produces similar results (data not shown).

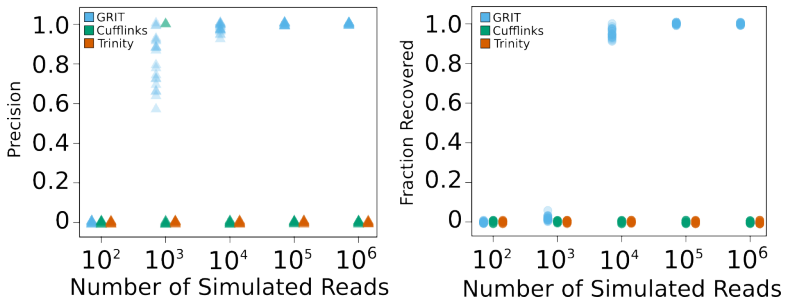

FIGURE 5.3.4. **Simulations - (a) *Dscam1* Simulations:** We simulated from the 38016 potential transcripts identified in Flybase 5.45. Trinity was not able to reconstruct any full length transcripts; Cufflinks was only able to construct a single full length transcript in 1/100 simulations. GRIT recovered most transcripts with high average precision when provided 1000 reads, and was able to reconstruct all 38016 transcripts with perfect precision when provided at 10,000 or more reads.

5.4. Discussion

The development of tools that enable the accurate interpretation of RNA sequence data is an important challenge. Our tool, GRIT, can leverage multiple RNA sequence data types, including CAGE, mRNA-seq, polyA+seq, ESTs, and cDNAs to discover transcript models. The use of gene boundary data prevents fragmentary transcript models, or models that erroneously merge distinct genes.

Transcript models assembled by GRIT begin with a transcript start site, are connected by intervening mRNA-seq signal, and end in a polyadenylation site. When applied to a subset of the modENCODE Drosophila RNA data sets[4], GRIT performs substantially better than competing methods, both at identifying previously annotated transcript models and at discovering of new genes and transcripts. GRIT also uses a novel transcript quantification procedure which correctly accounts for model unidentifiability when estimating the confidence bounds, permitting conservative confidence bounds even in gene loci with the potential to produce thousands of transcript isoforms.

In cases where the extant set of transcripts cannot be confidently identified, GRIT could be coupled with other classes of genomic information, including conservation, protein functional data, and RNA structure to produce a sparse subset of transcripts that preserve known function. This may aid in generating high-quality transcript annotations. As third generation sequencing technologies mature, it may become possible to observe full-length transcripts directly. GRIT currently incorporates cDNA sequences into transcript models, providing valuable connectivity information, and will make use of single molecule data-types as they become available.

Among the most remarkable findings of our work on the modENCODE Drosophila RNA datasets is the fact that over 20% of genes encode proteins with alternative localization signals. The gene Ptth has been studied for decades, yet GRIT discovered a new start codon, modulated by an alternative promoter. In addition to emphasizing the importance of accurate gene boundary information, our studies make evident the need for well-resolved tissue and cell-type transcript maps: the isoform in question is expressed in only two of the 108 modENCODE samples, where it is the dominant form. Future functional studies are needed to determine the biological role of this protein and indeed of the thousands of newly predicted protein isoforms with previously undetected protein localization signals.

Next generation sequencing has provided a view of transcriptomes with unprecedented depth and enormous complexity. GRIT generates full-length transcript models with sample-by-sample expression scores. This tool alleviates a current analytical bottleneck, and will dramatically enhance the accessibility and usefulness of RNA sequencing data.

CHAPTER 6

Application to modENCODE

6.1. Preface

The tools described in previous chapters form the basis of a a comprehensive analysis set. Their largest application to date has been to the modENCODE Droshila RNA data set. Statmap was used to map the PAS-Seq and CAGE data, and GRIT was used to build transcript models and quantifcations.

Below I present the major results of our analysis. The work below formed the basis for a manuscript which, after modification and substantial edits, was submitted to the journal Nature where it appeared under the title "Diversity and dynamics of the Drosophila transcriptome".

6.2. Introduction

Next-generation RNA sequencing has permitted the mapping of transcribed regions of the genomes of a variety of organisms. These studies demonstrated that large fractions of metazoan genomes are transcribed and cataloged individual elements of transcriptomes, including promoters, polyadenylation sites, exons and introns. However, the complexity of the transcriptome arises from the combinatorial incorporation of these elements into mature transcript isoforms. Studies that inferred transcript isoforms from short read sequence data focused on a small subset of isoforms, filtered using stringent criteria. Studies using cDNA or EST data to infer transcript isoforms have not had sufficient sampling depth to explore the diversity of RNA products at the majority of genomic loci. While the human genome has been the focus of intensive manual annotation, analysis of strand-specific RNA-seq data from human cell lines reveals over 100,000 splice junctions not incorporated into transcript models. Thus, a large gap exists between genome annotations and the emerging transcriptomes observed in next-generation sequence data. In Drosophila, we previously described a non-strand-specific RNA-seq analysis of a developmental time course through the life cycle and CAGE analysis of the embryo, which discovered thousands of unannotated exons, introns and promoters, and expanded coverage of the genome by identified transcribed regions, but not necessarily transcript models. Here, we describe an expansive poly(A)+ transcript set modeled by integrative analysis of transcription start sites (CAGE and 5' RACE), splice

sites and exons (RNA-seq), and polyadenylation sites (3' ESTs, cDNAs and RNA-seq). We analyzed poly(A)+ RNA data from a diverse set of developmental stages, dissected organ systems and environmental perturbations, much of which is strand-specific. Our data provide higher spatiotemporal resolution and allow for deeper exploration of the Drosophila transcriptome than was previously possible. Our analysis reveals a transcriptome of high complexity that is expressed in discrete, tissue- and condition-specific mRNA and ncRNA transcript isoforms that span the majority of the genome and provides valuable insight into metazoan biology.

6.3. Results

6.3.1. A dense landscape of discrete poly(A)+ transcripts. To broadly sample the transcriptome, we performed strand-specific, paired-end sequencing of poly(A)+ RNA in biological duplicate from 29 dissected tissue samples including the nervous, digestive, reproductive, endocrine, epidermal, and muscle organ systems of larvae, pupae and adults. To detect RNAs not observed under standard conditions we sequenced poly(A)+ RNA in biological duplicate from 21 whole-animal samples treated with environmental perturbations. Adults were challenged with heat-shock, cold-shock, and exposure to heavy metals (cadmium, copper and zinc), the drug caffeine, and the herbicide paraquat. To determine if exposing larvae resulted in novel RNA expression we treated them with heavy metals, caffeine, ethanol and rotenone. Lastly, we sequenced RNA from 21 previously described and three ovary-derived cell lines (Supplementary Methods). In total, we produced 12.4B strand-specific read-pairs and over a terabase of sequence data, providing 44,000 fold coverage of the poly(A)+ transcriptome.

Reads were aligned to the Drosophila genome as described, and full-length transcript models were assembled using our custom pipeline, GRIT (Supplementary Methods). GRIT uses RNA-seq, p(A)+seq, CAGE, RACE, ESTs, and full-length cDNAs to generate gene and transcript models. We integrated these models with our own and community manual curation datasets to obtain an annotation (Supplementary Material, section 12) consisting of 304,788 transcripts and 17,564 genes (Fig. 1a and Supplementary Fig. 1), of which 14,692 are protein-coding (Supplementary Data File 1). Ninety percent of genes produce at most 10 transcript and five protein isoforms, while 1% of genes have highly complex patterns of alternative splicing, promoter usage, and polyadenylation, and may each be processed into hundreds of transcripts (Fig. 1a, example 1b). Our gene models span 72% of the euchromatin, an increase from 65% in FlyBase 5.12 (FB5.12), the reference annotation at the beginning of the project (Supplementary Table 1 compares annotations 2008-2013). There were 64 euchromatic gene-free regions longer than 50kb in FB5.12, and 25 remaining in

FB5.45. Our annotation includes new gene models in each of these regions. Newly identified genes (1468 total) are expressed in spatially- and temporally-restricted patterns (Supplementary Fig. 2), and 536 reside in previously uncharacterized gene-free regions. Others map to well-characterized regions, including the ovo locus, where we discovered a new ovary-specific, poly(A)+ transcript (Mgn94020), extending from the second promoter of ovo on the opposite strand and spanning 107kb (Fig. 1c). Exons of 36 new genes overlap molecularly defined mutations with associated phenotypes (GSC p-value~0.0002), suggesting potential functions (Supplementary Table 2). For instance, the lethal P-element insertions l(3)L3051 and l(3)L4111 map to promoters of Mgn095159 and Mgn95009, respectively, suggesting these may be essential genes. Nearly 60% of the intergenic transcription we reported is now incorporated into gene models.

6.3.2. Transcript Diversity. Over half of spliced genes (56%) encode two or more transcript isoforms with alternative first exons (AFEs). The majority of such genes produce AFEs through coordinated alternative splicing and promoter usage (59%, 4389 genes, hypergeometric p-value<1e-16), suggesting coordination between these processes; however a substantial number of genes utilize one, but not both mechanisms (Fig. 2a). Only 1058 spliced genes have AFEs that alter coding capacity and increase the complexity of the predicted proteome. Some genes, such as G protein β-subunit 13F (Gβ13F, Fig. 2b, Supplementary Fig. 3) have exceptionally complex 5'UTRs, but encode a single protein.

We measured splicing efficiency using the "percent spliced in" (Ψ) index – the fraction of isoforms that contain the exon. Introns flanked by coding sequence are retained at an average Ψ=0.7, whereas introns flanked by non-coding sequence are retained >5-fold more often, with an average Ψ=3.8 (p<1e-16 subsampling/2-sample t-test), and is most frequent in 5'UTRs (mean Ψ=5.1, Fig. 2c).

Despite the depth of our RNA-seq data, our data show that 42% of genes encode only a single transcript isoform, and 55% encode a single protein isoform (Supplementary Methods). In mammals, it has been estimated that 95% of genes produce multiple transcript isoforms, (estimates for protein-coding capacity have not been reported).

The majority of transcriptome complexity is attributable to forty-seven genes that have the capacity to encode >1000 transcript isoforms each (Supplementary Table 3), and account for 50% of all transcripts (Fig. 3a). Furthermore, 27% of transcripts encoded by these genes were detected exclusively in samples enriched for neuronal tissue, and another 56% only in the embryo (83% total). To determine their tissue specificities we conducted embryonic in situ expression assays (Fig. 3b) and found that 18 of 35 are detected only in neural tissue (51% vs. 10% genome-wide, hypergeometric p-value<1e-16, Supplementary Table 4). Of these

genes, 48% have 3'UTR extensions in embryonic neural tissue (5% genome-wide, p<1e-16). Furthermore, 44% are targets of RNA-editing (4% genome-wide, p<1e-16, with 18 of 21 validated), and 21% have 3'UTR extensions and RNA-editing sites (10 of 65 genome-wide, p<1e-100). The capacity to encode thousands of transcripts is largely specific to the nervous system and coincides with other classes of rare, neural-specific RNA processing.

6.3.3. Tissue- and sex-specific splicing. To examine the dynamics of splicing, we calculated switch scores, or $\Delta\ \Psi$, for each splicing event by comparing the maximal and minimal Ψ values across all samples, and in subsets including just the developmental and tissue samples (Fig 4a). In contrast to the median Ψ values, the distribution of $\Delta\ \Psi$ values is strikingly different between the developmental and tissue samples. Among the developmental samples, 38% of events have a $\Delta\ \Psi$ 50% while between the tissue samples 63% of events have a $\Delta\ \Psi$50%. This difference is even more pronounced at higher $\Delta\ \Psi$ thresholds only 6% of events have a $\Delta\Psi$ 80% between the developmental samples while 31% of events have a $\Delta\ \Psi$80% between the tissue samples. Thus, most splicing events are highly tissue-specific. Of the 17,447 alternative splicing events analyzed (Supplementary Materials, section 19), we find that 56.6% changed significantly ($\Delta\ \Psi$>20%, BF>20). Clustering revealed groups of splicing events that are coordinately regulated in a tissue-specific manner. For example, 1147 splicing events are specifically included in heads and excluded in testes or ovaries, while 797 splicing events are excluded in heads but included in testes or ovaries (Fig. 4a).

We identified hundreds of sex-specific splicing events from adult male and female RNA-seq data. To further explore sex-specific splicing, we compared the splicing patterns in male and female heads enriched for brain tissues. There were striking differences in gene expression levels, however, only seven splicing events were consistently differentially spliced at each time point after eclosion (average $\Delta\ \Psi$>20%), and these largely corresponded to genes in the known sex-determination pathway (Supplementary Material). We find few examples of head sex-specific splicing. This is in contrast to previous studies, which have come to conflicting conclusions and used either microarrays analyzing only a subset of splicing events or single read 36bp RNA-Seq with an order of magnitude fewer reads.

We identified 575 alternative splicing events that are differentially spliced in whole male and female animals ($\Delta\ \Psi$>20%) and analyzed the tissue-specific splicing patterns of each event (Fig. 4b). We found that 186 of the 321 male-biased splicing events were most strongly included in testes or accessory glands, and 157 of 254 female-biased exons were ovary-enriched. Consistent with the extensive transcriptional differences observed in testes compared to other tissues, the genes containing male-specific exons are enriched in functions related to transcription. In contrast, the female-specific exon containing genes are enriched

in functions involved in signaling and splicing (http://reactome.org, Supplementary Table 6). Together, these results indicate that the majority of sex-specific splicing is due to tissue-specific splicing in tissues present only in males or females.

6.3.4. Long non-coding RNAs. A growing set of candidate long non-coding RNAs (lncRNAs) have been identified in Drosophila. In FB5.45 there were 392 annotated lncRNAs, and it has been suggested that as many as 1119 lncRNAs may be transcribed in the fly. However, this number was based on transcribed regions, not transcript models, and utilized non-stranded RNA-seq data. We find 3880 genes produce transcripts with ORFs encoding fewer than 100 amino acids (aa). Of these, 795 encode conserved proteins (Methods) longer than 20aa. For example, a single exon gene in the last intron of the early developmental growth factor spätzle encodes a 42aa protein that is highly conserved across all sequenced Drosophila species. We identified 1875 candidate lncRNA genes producing 3085 transcripts, 2990 of which have no overlap with protein-coding genes on the same strand (Supplementary Data File 2). Some of these putative lncRNAs may encode short polypeptides, e.g. the gene tarsal-less encodes three 11aa ORFs with important developmental functions. We determined protein conservation scores for each ORF between 20 and 100aa (Supplementary Table 6). Of the 1119 predicted lncRNAs, we provide full-length transcript models for 246 transcribed loci; the remainder were expressed at levels beneath thresholds used in this study. This is not surprising, the expression patterns of lncRNAs are more restricted than those of protein-coding genes: the average lncRNA is expressed (BPKM >1) in 1.5 developmental and 3.2 tissue samples, compared to 6.6 and 17 for protein-coding genes, respectively. Many lncRNAs (563 or 30%) have peak expression in testes, and 125 are detectable only in testes. Similarly restricted expression patterns have been reported for lncRNAs in humans and other mammals.

Interestingly, all newly annotated genes overlapping molecularly defined mutations with phenotypes are lncRNAs (Supplementary Table 2). For instance, the mutation D114.3 is a regulatory allele of spineless (ss) that maps 4 kb upstream of ss and within the promoter of Mgn4221. Similarly, Mgn00541 corresponds to a described, but not annotated 2.0 kb transcript overlapping the regulatory mutant allele ci[57] of cubitus interruptus. It remains to be determined whether these mutations are a result of the loss of function of newly annotated transcripts or cis-acting regulatory elements (e.g. enhancers) or both.

6.3.5. Antisense transcription. Drosophila antisense transcription has been reported, but the catalog of antisense transcription has been largely limited to mRNA-mRNA overlaps. We identify non-coding antisense transcript models for 402 lncRNA loci that are antisense to mRNA transcripts of 422 protein-coding genes (e.g. prd, Fig. 5a), and 36 lncRNAs form

"sense-antisense gene-chains" overlapping more than one protein-coding locus, as observed in mammals. In Drosophila, 21% of lncRNAs are antisense to mRNAs, whereas in human 15% of annotated lncRNAs are antisense to mRNAs (GENCODE v10). We assembled antisense transcript models for 5057 genes (29%, compared to previous estimates of 15%). For 67% of these loci, antisense expression is observable in at least one cell line, indicating that sense/antisense transcripts may be present in the same cells. LncRNA-mediated antisense accounts for a small minority of antisense transcription – 94% of antisense loci correspond to overlapping protein-coding mRNAs transcribed on opposite strands, and of these, 323 loci (667 genes) share overlapping CDSs. The majority of antisense is due to overlapping UTRs: 1389 genes have overlapping 5'UTRs (divergent transcription), 3430 have overlapping 3'UTRs (convergent transcription), and 540 have both, meaning that, as with many lncRNAs, they form gene-chains across contiguously transcribed regions. A subset of antisense gene-pairs overlap almost completely (>90%), which we term reciprocal transcription. There are 13 such loci (Supplementary Fig. 5) and seven are male-specific (none are female-specific).

The mRNA/lncRNA sense-antisense pairs tend to be more positively correlated in their expression than mRNA/mRNA pairs, (mean r~0.16 vs. 0.13, KS 2-sample one-sided test p<1e-9), and while this mean effect is subtle, the trend is clearly visible in the quantiles (95^{th}% lnc/mRNA 0.729 vs. m/mRNA 0.634, Supplementary Fig. 6a). This effect is stronger when the analysis is restricted to cell line samples (Supplementary Fig. 6b).

Even in homogenous cell cultures, evidence for sense-antisense transcription does not guarantee that both transcripts exist within individual cells: transcription could originate from exclusive events occurring in different cells. Cis-natural antisense transcripts (cis-NATs) are a substantial source of endogenous siRNAs, and their existence directly reflects the existence of precursor dsRNA. Cis-NAT-siRNA production typically involves convergent transcription units that overlap on their 3' ends, but other documented loci generate siRNAs across internal exons, introns or 5'UTRs. Analysis of head, ovary and testis RNAs showed that 328 unique sense/antisense gene pair regions generated 21nt RNAs indicative of siRNA production (Supplementary Table 8), and these were significantly enriched (Supplementary Figure 7a, Supplementary Methods) for pairs showing positively correlated expression between sense and antisense levels across tissues (p~2e-5), embryo developmental stages (p~4e-3), conditions (p~9e-4), and across all samples (p~3e-5). The tissue distribution of these cis-NAT-siRNAs showed a bias for testis expression (Supplementary Fig. 7b), with 4-fold greater number relative to ovaries (p~2e-17, binomial test) and 7-fold relative to heads (p~4e-25) and expression levels of siRNAs were substantially higher in testes than other tissues (Supplementary Fig. 7c).

Over 80% of cis-NAT-siRNAs were derived from 3'-convergent gene pairs. Abundant siRNAs emanate from an overlap of the gryzun and CG14967 3'UTRs (Supplementary Fig. 5). The remainders were distributed amongst CDSs, introns, and 5'UTRs. We identified abundant, testis-enriched, siRNA production from a 5'-divergent overlap of Cyt-c-d and CG31808 (Fig. 5b) and from the entire CDS of dUTPase and its antisense noncoding transcript Mgn99994.

6.3.6. Environmental stress reveals new genes, transcripts and common response pathways. Whole-animal perturbations each exhibited condition-specific effects, e.g. the metallothionein genes were induced by heavy metals (Fig. 6a), but not by other treatments (Supplementary Table 9). The genome-wide transcriptional response to cadmium (Cd) exposure involves small changes in expression level at thousands of genes (48 hours after exposure), but only a small group of genes change >20-fold, and this group includes six lncRNAs (the third most strongly induced gene is CR44138, Fig. 6a, Supplementary Fig. 8a). Four newly modeled lncRNAs are differentially expressed (1% FDR) in at least one treatment, and constitute novel eco-responsive genes. Furthermore, 57 genes and 5259 transcripts (of 811 genes) were detected exclusively in these treatment samples. Although no two perturbations revealed identical transcriptional landscapes, we find a homogeneous response to environmental stressors (Fig. 6b, Supplementary Fig. 8b). The direction of regulation for most genes is consistent across all treatments; very few are up-regulated in one condition and down-regulated in another. Classes of strongly up-regulated genes included those annotated with the GO term "Response to Stimulus, GO:0050896" (most enriched, p-value<1e-16, Supplementary Fig. 8c), and those that encode lysozymes (>10-fold), cytochrome P450s, and mitochrondrial components mt:ATPase6, mt:CoI, mt:CoIII (>5-fold). Genes encoding egg-shell, yolk, and seminal fluid proteins are strongly down-regulated in response to every treatment except "Cold2" and "Heat Shock" (Supplementary Fig. 8d). For these two stressors, samples were collected 30 minutes after exposure, corresponding to an "early response test" showing suppression of germ cell production is not immediate.

6.4. Discussion

The majority of transcriptional complexity in Drosophila occurs in tissues of the nervous system, and particularly in the functionally differentiating central and peripheral nervous systems. A subset of ultra-complex genes encodes more than half of detected transcript isoforms and these are dramatically enriched for RNA-editing events and 3'UTR extensions, both phenomena largely specific to the nervous system. Our study indicates that the total information output of an animal transcriptome may be heavily weighted by the needs of the developing nervous system.

The improved depth of sampling and spatiotemporal resolution resulted in the identification of more than 1200 new genes not discovered in our previous study of Drosophila development. A large fraction of the new genes are testes-specific, and many of these are antisense RNAs, as previously described in mammals. Some new lncRNAs, such as Mgn94020 (Fig. 1), form sense/antisense gene-chains that bring distant protein-coding genes into transcriptional relationships, another phenomenon previously described only in mammals. Whenever Mgn94020 is detectably transcribed, the genes on the opposite strand in its introns are not, suggesting that its transcription may serve a regulatory function independent of the RNA transcribed. The presence of short RNAs at many regions of antisense transcription indicates that sense and antisense transcripts are present in the same cells at the same times. Many of these Drosophila antisense transcripts correspond to "positionally equivalent" antisense transcripts in human. In the two species we found antisense lncRNAs opposite to orthologous protein-coding genes. The apparent positional equivalence of fly and human antisense transcription at genes like Monocarboxylate transporter 1 (Mct1), even-skipped (EVX1), CTCF (CTCF), Adenosine receptor (ADORA2A), and many others across 600 million years of evolution suggests a conserved regulatory mechanism basal to sexual reproduction in metazoans.

Perturbation experiments identified new genes and transcripts, but perhaps more importantly, a general response to stress that is broader than the heat shock pathway. A similar study conducted on marsh fishes in the wake of the Deep Water Horizon incident in the Gulf of Mexico demonstrated that the killifish response to chronic hydrocarbon exposure included induction of lyzosome genes, P450 cytochromes, and mitochondrial components, and the down-regulation of genes encoding egg-shell and yolk proteins. This overlap of expressional responses by gene families across phyla suggests a conserved metazoan stress response involving enhanced metabolism and the suppression of genes involved in reproduction.

We defined an extensive catalog of putative lncRNAs. However, many genes are known to encode poorly conserved, short polypeptides, including genes specific to the male gonad and accessory gland. Ribosome profiling (Ribo-seq) initially indicated that a number of putative mammalian lncRNAs may be translated, but this observation has been difficult to validate by proteomics, and a re-analysis of Ribo-seq data has suggested that lncRNAs although they have signatures of ribosome occupancy are not translated. Therefore, while we refer to these RNAs as "non-coding", additional data are needed to determine if they produce small polypeptides.

Our observations raise many questions. Why do genes encoding RNA binding proteins exhibit extraordinary splicing complexity, often within their 5'UTRs? The splicing factor pUf68 encodes more than 100 alternatively spliced 5'UTR variants, but encodes a single protein. The notion that splicing factors may regulate one another to generate complex

patterns of splicing is consistent with recent computational models. What is the role of complex splicing during the development of the nervous system? To answer the questions that come with increasingly complete transcriptomes in higher organisms, it will be necessary to study gene regulation downstream of transcription initiation, including the regulation of splicing, localization and translation.

6.5. Figures

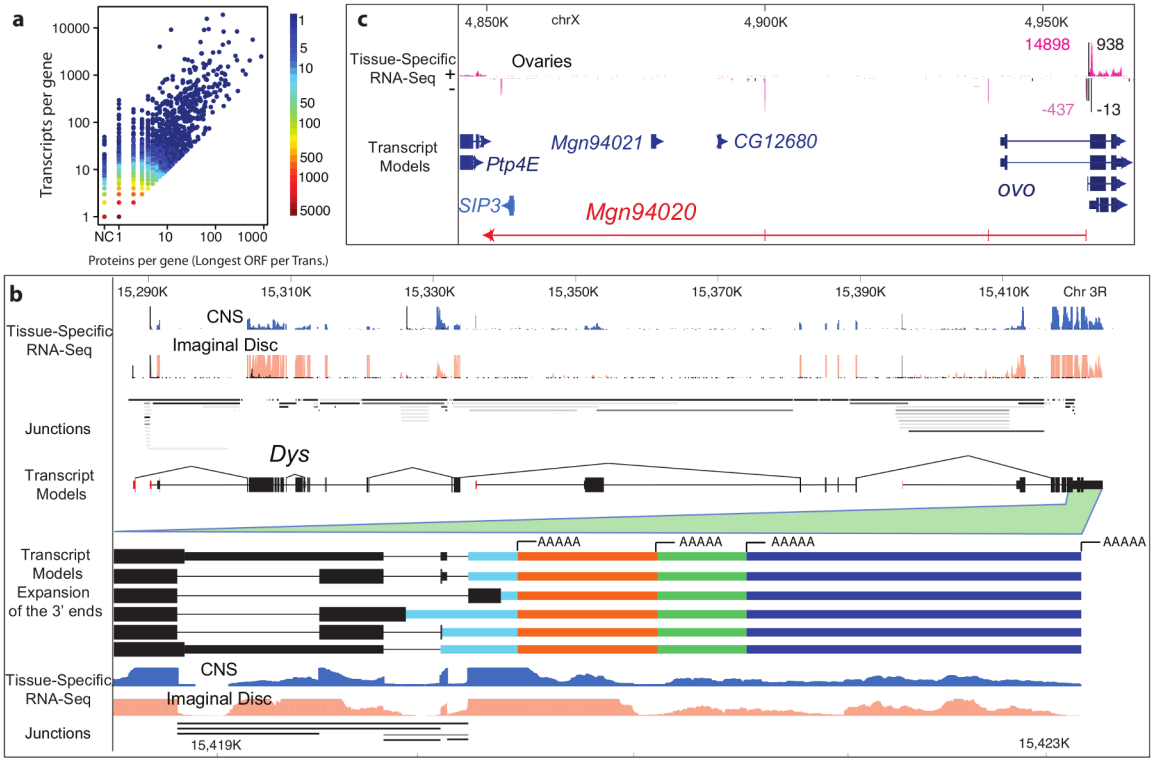

FIGURE 6.5.1. Overview of the annotation a, Scatterplot showing the per gene correlation between number of proteins and number of transcripts. The genes Dscam and para are omitted as extreme outliers both encoding >10,000 unique proteins. b, Dystrophin (Dys) produces 72 transcripts and encodes 32 proteins. Highlighted is alternative splicing and polyadenylation at the 3' end. c, An internal promoter of ovo is bidirectional in ovaries and produces a lncRNA (430bp) bridging two gene deserts.

6.6. Supplementary Methods and Results

6.6.1. Fly rearing and developmental staging. Fly stocks (except where specified, the sequenced D. melanogaster isogenic strain y^1 cn^1 bw^1 sp^1 was used[1]) were reared at 24 C on standard Drosophila medium. To collect larvae and adults, the flies were raised in

250 ml bottles containing 40 ml medium. To aid in staging third instar larvae the medium contained 0.05% bromphenol blue (BPB[2]) and staging was done as described[3].

Synchronized embryos were collected from large population cages (ca. 25 cm x 25 cm x 25 cm; maintained at 24 C on a cycle of 14 h light 10 h dark) from adults that were less than one week old. Following at least one – 2 h pre-lay that preceded timed collections each day, embryos were collected for two hours on three hard egg lay collection plates made in 150 X 15 mm Petri dishes containing a substrate of 3.3% agar, 13% unsulfured molasses, and 0.15% Tegasept. The hard egg lay plates were completely covered with a thin layer of moist yeast paste (Fleischmann's Baker's Dry Yeast) and placed horizontally on a short 1 cm raised Plexiglas bar in the bottom of each cage to avoid crushing flies. Staged embryos were passed through an 850 micron screen and collected on a 75 micron screen to remove adults and yeast paste. Embryos were then dechorionated by treatment with a solution of 50% bleach (3% sodium hypochlorite), 0.2% sodium chloride, and 0.02% Triton-X-100 for five minutes. Embryos were washed twice with 0.2% NaCl, 0.02% Triton buffer and split into two samples. Most of the sample (approximately 95%) was rinsed with de-ionized water in a buchner funnel under mild vacuum, dried briefly, immediately frozen on dry ice and stored at -80 C for RNA preparations. The small aliquot was transferred to a clean tube and fixed (0.1 M Pipes (pH 6.9), 2 mM EGTA, 1 mM MgSO4, 4% paraformaldehyde, 0.1% glutaraldehyde and 50% heptane for staging[4]. Samples were shaken for five minutes in the fixative, centrifuged briefly and the aqueous fraction was removed. An equal volume of methanol containing 2 mM EGTA was added and the sample was shaken for five additional minutes. Tissue was washed twice in methanol with 2 mM EGTA and saved at -80C for the characterization of developmental stages.

6.6.2. Dissection of Organ Systems. To detect rarely expressed and tissue specific RNAs we dissected organ systems from larval, pupal and adult animals. We examined components of the nervous system, from larval and pupal brains and ventral ganglia and from aged 1, 4 and 20-day adult heads (primarily brain) of mated males and virgin and mated females. To interrogate the reproductive system we dissected ovaries from females and testes and accessory glands from males. To study the digestive system we examined larval and pupal salivary glands and larval and aged 1, 4 and 20-day adult midgut, hindgut and malpighian tubules. We dissected larval and pupal fat body the primary metabolic and detoxification organ performing functions analogous to the human liver. To study the epidermis and muscle organ systems, we mass isolated larval imaginal discs adapted from a previously describe approach[2], with modifications detailed below and an aliquot of the sample prep is shown in Supplementary Figure 8. We also dissected larval and aged 1,4 and 20-day adult carcasses, which contain cuticle, epidermis, muscle and oenocytes as well as peripheral

neurons. All tissues were stored at -80 C immediately after dissection until sufficient material had been collected to permit RNA preparations. A yield of approximately 4 μg total RNA per mg of tissue collected was typical. A cartoon giving the anatomical relationships between the tissues collected is provided in Supp. Fig. 10. Specifics follow:

2A. Larval tissue dissections: Bottles were started with approximately 60 adult OreR flies at 25 C. After 5 days, climbing third instar larvae were collected and transferred to a dissecting surface with 1X PBS buffer (Ambion) for dissections. We identified the sex of the larva by the presence of the large clear spherical testes (or smaller ovary) embedded in the white fat body on the lateral sides of the A5 segment. We recorded and collected the tissues with equal representation of each sex. To dissect, the cuticle was torn immediately posterior to the mouth hooks using paired forceps and the larvae were everted as with WPP dissections. The digestive system and fat body were pulled toward the anterior end and away from the cuticle. The digestive system was disconnected from the body immediately anterior to the proventriculus. The salivary glands were collected by pinching them off from the attached fat body. The extensive and reticulated fat body was removed from the carcass and digestive system. The trachea were removed from the digestive system and collected with the carcass. Tissues collected included the gut (fat body removed, Malphigian tubules included), the salivary glands (with as much fat body removed as possible), and carcass (without the guts, salivary glands, fat body and gonads). Dissections were done concurrently so that all three tissues were collected from a single animal. Male and female tissues were collected in separate tubes and mixed in equal numbers for the RNA preparations.

2B. L3 Imaginal Discs mass preparation: Bulk preparations of imaginal disc tissue were done as previously described[5] with the following modifications. First instar larvae were transferred to ventilated plastic chambers containing seventeen feet of cotton rope saturated in a protein-rich yeast slurry (200 g active dry yeast, 6 oz Gerber's Banana food, 100 ml Grapefruit juice, 50 g ground Special K, 40 g Gerber's Baby Cereal, 20 g Wheat Germ, 1200-1400 ml water) and were allowed to grow until wandering larvae were observed. Larvae were ground with a Kitchen Aid Artisan mixer (Model KSM150SPER) and Kitchen Aid grain mill attachment (Model KGMA) with the plates set to leave about 5% of the total larvae unground. Ringer's solution was replaced with Organ Medium (25 mM β-Glycerol phosphate disodium salt pentahydrate (Fluka 50020), 10 mM KH_2PO_4, 30 mM KCl, 10 mM $MgCl_2$, 3 mM $CaCl_2$, 162 mM sucrose) at all steps. A photograph of the isolated tissues is given (Supplementary Fig. 9).

2C. Fly WPP and 2-day old pupae CNS: Staged WPP and 2-day old pupae were dissected in PBS (phosphate buffered saline). The posterior end of the pupa was removed with two forceps at the A7 abdominal segment. The anterior body of the pupa was removed from

the pupal case with forceps. We held each cuticle at the anterior tip and gently teased the body towards the posterior opening with forceps. We pulled the cuticle from the anterior end through the second forceps, holding them nearly closed around the vacated cuticle. This squeezed the body of the pupa out of the cuticle. The yellow eye discs were removed from the brain lobes of the CNS. The connected antennal segment at the anterior margin of the brain was removed. The developing leg and wing disc tissue along with the fat body was removed, and the attached subesophageal ganglion and ring gland were recovered along with the brain. The CNS and ring gland were transferred to a collection tube on dry ice and then stored at -80 C until sufficient tissue for RNA isolation was collected.

2D. White pre-pupal salivary gland and fat body: We collected white pre-pupae (WPP) as in Graveley et al. [3] and dissected in PBS buffer. We identified the sex of the larva by the presence of the large clear spherical testes (or smaller ovary) embedded in the white fat body on the lateral sides of the A5 segment. We recorded and collected the tissues with equal representation of each sex. We note that the female WPP tend to be larger. We tore the cuticle immediately posterior to the mouth hooks and then everted the WPP by pushing the posterior end inside the body cavity with closed forceps, and finally collected the fat body and salivary glands in separate tubes on dry ice.

2E. Pupal fat body mass preparation: We transferred WPP animals 48 h after staging and resting at 25 C to a 15 ml polycarbonate falcon tube. We added 2 ml of Drosophila Ringers (182 mM KCl, 46 mM NaCl, 3 mM CaCl, 10 mM Tris-HCl pH 7.2) containing 2% Ficoll, and crushed the pupae in the tube to release contents from the cuticles. We added 5 ml of Ringers with 2% Ficoll, mixed with a large bore disposable pipet and filtered through a 100 um screen. The cell suspension was centrifuged at 660xG for 10 minutes at 40C, and fat body cells were collected from the surface of the buffer and transferred to a 1.5ml eppendorf tube. Cell suspensions were centrifuged at 660xG for 5 minutes at 40 C to remove as much of the buffer from beneath the cells as possible. We froze the fat body cells by placing them on dry ice and stored at -80 C for RNA preparation.

2F. Adult gonads and reproductive tissues: Staged adult flies were anesthetized with CO_2 for 30 minutes or less while dissections were done in PBS (1X; Ambion) for less than 10 minutes each. To dissect/open the abdomen, we pinned down the thorax on either side with a set of surgical steel forceps (size #4 or #5), and pulled the T3 legs posteriorly to remove the overlying cuticle and expose the digestive and reproductive organs. The reproductive tissues were removed and separated from the digestive tract and the cuticle. In the females, the reproductive tissues included the ovaries and their attached oviducts. Due to tearing and mechanical damage during dissection, the oviducts were incompletely recovered. In the males, the reproductive tissues included the testes (generally bright yellow), and the

accessory glands (generally translucent, with incomplete recovery of the attached seminal vesicle). We collected the ovaries and oviducts together as a single sample, and separated the testes from the accessory glands for independent RNA isolation and sequencing. These tissues were dissected away from all other attached cells, and then frozen in 1.5 ml tubes submerged in dry ice, and then stored at -80 C until sufficient quantities were obtained for RNA purifications.

2G. Adult gut and carcass: Adult flies were staged and anesthetized as for the gonad preparation. The digestive tracts and carcasses were separated after removing the head and discarding. Holding the thorax and pulling the T3 legs posteriorly to expose the digestive and reproductive organs was used to dissect the abdomen. The reproductive tissues were removed and discarded. The digestive tract was separated from the cuticle, adipose tissue (fat body) and other tissues, and then frozen on dry ice. The remaining tissues (without the head and reproductive organs) were frozen on dry ice and designated the carcass. All tissues were stored at -80 C until sufficient quantities were obtained for RNA purifications.

2H. Adult head: Isolation of the fly heads was accomplished by placing CO_2-anesthitized adults in a 15 ml conical tube that was then flash frozen in liquid nitrogen for about one minute. The tube was then shaken vigorously for 10 seconds, and tapped on the bench-top. The broken flies were placed in a frozen glass petri dish on dry ice. The frozen severed heads were removed with dissecting forceps and placed in an eppendorf tube on dry ice. Flies were processed in groups of 100 animals per dissection. Isolated tissue was stored at -80 C until RNA could be purified from an adequate number of prepared heads. Typically, heads were missing the antennal and maxillary organs, while the mouth-parts were retained.

6.6.3. Environmental Perturbations. 3A. Heat Shock: Twenty virgin males and 20 virgin females were maintained on standard corn meal agar at 25Â°C for four days. After four days the 40 adult flies were transferred to clean glass vials and placed in a 36Â° C water bath (wet heat) and held at 36Â°C for 1 hour followed by a 30-minute recovery at 25Â° C prior to freezing in liquid nitrogen. This treatment produced relatively high lethality due to excessive moisture buildup in the vials.

3B. Cold Shock1: Newly eclosed flies were collected, and placed in cornmeal agar food vials containing 20 males and 20 females were and kept at 25~ C for 84 hours. Aged, mated flies were transferred to empty glass vials and placed in a micro-cooler water bath containing 10% glycol at 25 C. The temperature was decreased to 0C at a rate of 0.2C per minute and then flies were held at 0 C for 9 hours. After the cold treatment flies were transferred to fresh food vials and kept at 25 C for 2 hours for the recovery period. Following recovery flies

were placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations.

3C. Cold Shock 2: Flies were treated as in "Cold Shock 1", above, except flies were held on food vials for four days. Aged, mated flies were transferred to empty glass vials and placed in a micro-cooler water bath containing 10% glycol at 0Â° C for two hours. Following the cold shock flies were transferred to fresh food vials and kept at 25° C for 30 minutes for the recovery period. Following recovery flies were placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80C prior to RNA preparations.

3D. Feeding schedule for consumed treatments:

3D1. Treatment schedule for Larvae: For each treatment, approximately 50 (mixed sex) young mated adults were transferred to each fresh food vials and maintained for 12 hours. Vials were cleared and allowed to age 3.5 to 4 days. Vials were then rinsed into a series of sieves using tepid water; feeding third instar larvae were collected form the #40 sieve and transferred to a hard agar plate with a pot of yeast to induce crawling. Prior to reaching the yeast, larvae were captured and 50 larvae were transferred to new food vials containing the treatment of interest (details below), and larvae were allowed to feed for 4 hours. Treated larvae were captured and transferred to 2 ml vials, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. The number of survivors was recorded and the mean lethality calculated for each treatment.

3D2. Treatment schedule for Adults: For each treatment, 40 newly eclosed males and females (1:1) were transferred to fresh food (BDSC corn meal agar) vials and maintained at 25 C for two days. To treat flies, two Kimwipes were folded into a square and put in the bottom of a one-pint glass bottle. Kimwipes were saturated with 4 ml of the treatment solution, (10% sucrose solution and one drop of green vegetable coloring per 50 ml solution, plus the treatment of interest). Harvesting time for adults varied by treatment. Upon harvesting, flies were placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations.

3D3. Caffeine feeding: Starved larvae (as above) were transferred to food vials containing 1.5 mg/ml caffeine and allowed to feed for 4 h. Adults fed 25 mg/ml caffeine were harvested after 8 h; adults fed 2.5 mg/ml caffeine were harvested after 48 h, and after 24 h an additional 1 ml of treatment solution was dripped onto the Kimwipe. Upon harvesting, flies were placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. For adults, 2.5 mg/ml caffeine is near the LD50 for a 48 h treatment. 25 mg/ml caffeine is 100% lethal after 24 h.

3D4. Copper feeding: Starved larvae were transferred to new food vials containing 0.5 mM $CuSO_4$ and allowed to feed for 12 h. The number of survivors was recorded and the mean lethality calculated for each treatment. Adults were fed with 15 mM $CuSO_4$. After 24 h an additional 1 ml of the treatment solution was dripped onto the Kimwipe. Flies were harvested after 48 h of feeding, placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. Adult concentrations were all done at or near the LD50 determined for our feeding method after 48 h. Adults were fed 15 mM copper for 48 h.

3D5. Zinc feeding: Starved larvae were transferred to new food vials containing the 5 mM $ZnCl_2$ and allowed to feed for 12 h. Treated larvae were transferred to 2 ml vials, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. Adults were fed with 4.5 mM $ZnCl_2$. After 24 h an additional 1 ml of the treatment solution was dripped onto the Kimwipe. Flies were harvested after 48 h of feeding, placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. Adult concentrations were done at or near the LD50 determined for our feeding method after 48 h. Adults were fed 4.5 mM zinc for 48 h. Zinc appears to cause a neuromuscular defect in both adults and larvae.

3D6. Cadmium feeding: Starved larvae were transferred to new food vials containing 0.05 mM $CdCl_2$ and allowed to feed for 6 or 12 h. Treated larvae were transferred to 2 ml vials, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. Adults were fed with 0.1 mM or 0.05 mM $CdCl_2$. After 24 h an additional 1 ml of solution was dripped onto the Kimwipe. Flies were harvested after 48 h of feeding, placed in 2 ml tubes, flash frozen in liquid nitrogen and stored at -80 C prior to RNA preparations. Adult concentrations were all done at or near the LD50 determined for our feeding method after 48 h. This concentration had a minimal effect on larvae after 6 h. Additionally, two vials of larvae were allowed to complete development and 96% eclosed with no obvious phenotypic abnormalities.

3D7. Paraquat feeding: Two-day-old adults were fed 5 mM paraquat for 48 h, and 3-day-old adults were fed 10 mM paraquat for 24 h. Following the treatment, adult flies were flash-frozen in liquid nitrogen and stored at -80 C. Feeding third-instar larvae were transferred to food containing 10 mM paraquat and allowed to feed for 12 h. Following treatment, larvae were collected and flash-frozen in liquid nitrogen and stored at -80 C.

3D8. Rotenone feeding: Newly eclosed adults were fed 20 μg/ml rotenone in 10% sucrose continuously for 10 days. Following the treatment adult flies were flash-frozen in liquid nitrogen and stored at -80 C. There was no evidence that the adults actually ingested any of the rotenone/sucrose/green dye solution, so we believe that any effect on transcription was likely to be caused by starvation rather than by rotenone itself. Hence we did not sequence

RNA from these flies. Feeding third-instar larvae were transferred to food containing either 2 μg/ml or 8 μg/ml rotenone and allowed to feed for 6 h. Following treatment, larvae were collected and flash frozen in liquid nitrogen and stored at -80 C.

3D9. Resveratrol feeding: Two-day-old adults were fed 100 μM resveratrol in 10% sucrose continuously and samples were harvested at 10 days. Adult flies were flash frozen in liquid nitrogen and stored at -80 C.

6.6.4. RNA isolation. RNA from whole animals and cell lines was isolated as previously described[3]. Tissues and organ system samples were homogenized in TRIzol reagent: the sample volume not to exceed 10% of the volume of TRIzol reagent, incubated at room temperature for 5 minutes before centrifugation in 1.5 ml microcentifuge tubes. Chloroform was added using 0.267 ml per ml of TRIzol, the tubes were mixed vigorously for 15 seconds, and incubate at room temperature for 2 minutes. Samples were centrifuged for 15 minutes at 4 C at 12,000g. The top (aqueous) phases were transferred to clean tubes. RNA was precipitated from the aqueous phase by adding 0.67 ml of isopropanol per ml of TRIzol. Tubes were inverted once to mix components. Samples were incubated at room temperature for 10 minutes and then centrifuged for 10 minutes at 4 C at 12,000g. The supernatant was removed and the RNA pellet washed once with 75% ethanol, using 0.7 ml per microcentrifuge tube with a brief vortex. We centrifuge at 7,500g for 5 minutes at 4 C and then let the pellet air dry for 10 minutes but did not dry completely. We dissolved the pellet in RNase-free water and incubated at 37C overnight to dissolve the RNA. The concentration of RNA was determined using a Nanodrop ND-1000 Spectrophotometer. RNA was stored at -80C for shipping purposes.

In addition we isolated RNA using the RNeasy (Qiagen) kit that does not capture the RNAs <200 nt. Poly(A)+ RNA-seq and CAGE were performed using RNeasy samples and thus reflect transcripts >200 nt.

6.6.5. Illumina RNA-seq library construction and sequencing. We performed stranded paired-end RNA sequencing using the Illumina TruSeq stranded sample preparation kit (Catalog No.15031048). The non-strand-specific RNA-Seq data from the developmental samples were previously described[3]. Strand-specific RNA-seq libraries were prepared from the tissue, cell line, and environmental samples using prerelease Directional mRNA-seq Library Kits (Illumina) as described previously[6]. Strand-specific total RNA libraries were prepared from the developmental RNA samples using the dUTP-based protocol described in[7]. The poly(A) enrichment libraries were prepared from the 29 tissue sample in biological

duplicate as described in[6]. Libraries were sequenced on the Illumina GAIIx or HiSeq2000 platforms using single or paired-end 76-100 bp chemistry.

6.6.6. Illumina CAGE library construction and sequencing. CAGE libraries were constructed from 36 total RNA samples (RNeasy, Qiagen) using the procedure described in[8]. The libraries were sequenced on the Illumina GAIIx platform to generate 36-nt reads. The 9-nt barcode linker sequence was removed, and the 27-nt CAGE reads representing capped 5' transcript ends were aligned to the D. melanogaster genome using StatMap.

6.6.7. RNA sequencing of polyadenylation sites. RNA sequencing libraries specific for polyadenylation sites were prepared as follows. RNA samples, from dissected heads of males and mated females at 20 days post-eclosion, were used to produce two "polyA-seq" libraries. Total RNA (2 g) was fragmented in 1X RNA Fragmentation Reagent (Life Technologies) in 10 l at 65Â°C for 5 minutes. The reaction was stopped by addition of 1 l of reaction stop buffer (Life Technologies) and cooled on ice. The fragmented RNA sample was used, without precipitation, as the starting material for the library construction protocol and kit described in the Illumina TruSeq Stranded mRNA Sample Preparation Guide (Rev. D, September 2012), with the following modifications. At the second round of poly(A)+ RNA selection, the bound RNA was eluted with addition of 13.5 l of nuclease-free water and heating to 65Â° for 5 minutes. The eluted RNA was removed from the beads in 11, and 1 l of a custom anchored oligo-dT primer (20 g/l; 5'-NGCAGCAT(20)VN-3') and 5 l of 5X Superscript II Buffer (Life Technologies) were added. The sample was heated to 42C for 2 minutes to anneal the primer, then cooled on ice.The annealed sample was used to prepare a sequencing library following the remaining steps in the Illumina protocol from first-strand cDNA synthesis to the end. Libraries were sequenced on the Illumina HiSeq platform to produced paired-end reads (2 x 100 nt) following standard protocols.

6.6.8. 454 Titanium-platform RNA-seq library construction and sequencing. Primer annealing and first strand synthesis was a modification of the Clonetech SMART protocol and used Superscript II from Invitrogen: 420 ng of RNA in water was used in first strand synthesis. To this was added 2 μl and 10 μM Cap-Tail primer at 65Â°C for 3 minutes and then on ice for 1 minute. To this was added 4 μl Clonetech 5X First Strand Buffer, 1 μl 10mM dNTP mix, 2 μl 0.1M DTT, 2ul 10 μM Clonetech Template-Switch Primer, 2 μl Superscript II reverse transcriptase and this was incubated at 42Â° C for 1.25 hours, and then at 70Â° C for 15 minutes and on ice for 2 minutes. Second strand cDNA synthesis and amplification used Quanta Biosciences AccuStart Polymerase and 16 cycles of amplification as follows. A solution containing 330 μl of water was combined with 42.5 10X buffer, 17 μl

50mM $MgSO_4$, 8.5 μl 10 mM dNTP, 17 μl 10 μM CAP Primer, 1.7 μl DNA Polymerase, and 8.5 μl first strand cDNAs (generated as above). This mix was divided into 4 aliquots of 100 μl each subjected to 16 cycles of: 94Â° C, 5 min; 94Â° C, 40 sec; 65Â° C, 1 min; 72Â° C, 4 min. Reactions were combined and cleaned using a single Qiagen QiaQuick PCR column, with elution in 30 μl Qiagen EB (10 mM Tris pH 8.0). This yielded cDNA at 710 ng/μl by Nanodrop spectrophotometry. Next, partial normalization of cDNA abundances was done using the Evrogen, Trimmer Direct Kit: double stranded nuclease (DSN) treatment for final library (1200 ng) was performed with 1/8 dilution of the DSN enzyme stock and 9 cycles of amplification. Next, the normalized cDNA library was divided into 6 aliquots of 100 μl each and amplified a further 9 cycles. Reactions were combined and cleaned using a single Qiagen QiaQuick PCR column, with elution in 30 μl Qiagen EB. Fragmentation to appropriate size for 454 sequencing was by nebulization: 400 ng cDNA was fragmented at 30 psi, 1 min. using a Roche Rapid nebulizer. Fragmented cDNA was concentrated using a single Qiagen Minelute column, with elution in 25 μl Qiagen EB. Fragments were end-polished and ligated to adaptors using reagents from the Roche GS-FLX Titanium General Library Preparation Kit, except for the fragmentation, which used the Klenow kit from New England BioLabs. To 375 ng of fragmented cDNA (9.4 μl) was added 1.5 μl 10x Polishing buffer, 1.5 μl BSA, 0.8 μl dNTP mix, 0.9 μl T4 DNA Pol, 0.9 μl Klenow fragment, which we incubated at 12Â° C, 15 min; 25Â° C, 15 min; 70Â° C, 15 min. Adapters were added per Titanium General Library kit instructions and the reaction was cleaned using a single Qiagen QiaQuick PCR column, with elution in 30 μl Qiagen EB. To selectively amplify properly ligated templates, suppression PCR was performed as follows: to 390 μl of water were added 52.5 μl of 10X buffer, 21 μl of 50 mM $MgSO_4$, 10.5 μl each of 10 μM Primers A and B, 5.3 μl of each of 0.5 μM Suppression Primers 1 and 2, 2 μl of DNA Polymerase and 17 μl of Ligation Products (as above). The mix was divided into 6 aliquots and subjected to 16 cycles of: 94Â° C, 5 min; 94Â° C, 40 sec; 65Â° C, 1 min; 72Â° C, 4 min. The reaction was cleaned using a single Qiagen QiaQuick PCR column, with elution in 20 μl Qiagen EB. Final size selection was by gel electrophoresis and solid phase reversible mobilization (SPRI) magnetic bead capture. Of this Library, 400 ng was combined with 400 ng of pre-fragmented library above and run at 100 V, 2 h on a 0.8% GTG SeaKem agarose/TAE gel with SybrSafe dye (Invitrogen). The fraction of templates corresponding to the 500 bp to 800 bp size range were excised and purified using the Qiagen QiaQuick Gel Isolation Kit according to the manufacturer with the exception that no heat was used to melt agaraose. The library was eluted in 50 μl Qiagen EB. The library was further size selected for removal of small fragments using 0.5X (25 μl) of AMPure (SPRI) beads according to the manufacturer (Agencourt), with elution in 20 μl Qiagen EB. Library is stored in a siliconized tube at -80 C.

6.6.9. Read mapping and filtering. RNA-seq reads were mapped as previously described[3]. RNA-seq reads mapping to splice junctions were filtered additionally using the GRIT pipeline under default parameters[11]. CAGE reads were mapped as in[9]. Long RNA-seq reads sequenced on the 454 Titanium platform were mapped using the Celniker cDNA mapping pipeline described in [3]. Reads ending in poly(A) signal from both paired-end Illumina RNA-seq and 454 Titanium-platform RNA-seq (1.84 M reads) were treated differently: we extracted all reads ending in at least 5 A's where the body of the read, but not the A's map to the genome uniquely (no more than 2 mismatches and one mapped site). This resulted in the identification of 111,158 potential polyadenylation site—s by at least one read, 9,161 of which were within known CDS exons with no prior evidence of internal polyadenylation events. Furthermore, 78% of these poly(A) sites lie more than 2 kb from known poly(A) sites in the genome, consistent with recently reports in human[7]. We note that these ubiquitous poly(A) events, however, constitute only a small fraction of all poly(A) reads: 80% of poly(A) reads were accounted for by known poly(A) sites (within 500 bp of the known site). Hence, we hypothesize that some background signal exists in either the bioinformatics (read mapping) or the biochemical assay, or both, that may lead to the appearance of either rare or artifactual polyadenylation events. To filter these, we trained a Random Forest classifier (RF) (sklearn version 8.7.1) using poly(A) reads within 50 bp of a poly(A) site confirmed by cDNA sequencing as true positives (Supplementary Data File 4), and poly(A) reads in annotated CDS exons and/or in intergenic or intronic space with no other RNA-seq reads within 500 bp as true negatives. We utilized local poly(A) read density, genome sequence and known poly(A) motifs in fly[12] as well as motifs obtained using MEME[13] on cDNA-confirmed poly(A) sites, and RNA-seq read density as covariates (for a list see Supplementary Data File 3). The fitted RF had sensitivity of 97% and an FDR of 3% under cross validation on a held-out test set. It should be noted that the purity of the negative control cannot be assured, and hence the true false positive rate may be much lower. We fitted the classifier 100 times with randomly selected test sets to compute the variability of the imputed sensitivity and FDR, and found both to have standard deviations of 1%. This process retained 57,594 poly(A) sites, accounting for 82% of all poly(A) reads and including 94 that remained within annotated CDS exons. We manually reviewed each of the 94 instances and in each case removed these polyadenylation events from our models. Hence, poly(A) reads lying near known poly(A) sites, or sites with similar sequence composition and patterns of RNA-seq coverage account for the vast majority of poly(A) reads. We note that our complete empirical poly(A) dataset is missing poly(A) sites for 757 genes, mostly low expression genes including gustatory, olfactory, and inotropic receptors. We manually reviewed each of these 757 loci. The majority had poly(A) ends from targeted cDNA sequencing from the literature, but others required manual annotation. When possible, we assigned 3' ends based on RNA-seq

coverage (first base with zero read coverage or 100 fold fall-off), otherwise we accepted the boundary assigned by FB5.45, which in some cases was a stop codon. Our complete 3' end annotation, including manual annotations, is given in Supplementary Data File 5.

6.6.10. Building transcript models from CAGE, RNA-seq, EST, cDNA, and poly(A) sequence data. We used the GRIT algorithm as described[11] with default parameters and the full set of our RNA-seq datasets to generate transcript models. To obtain sufficient sequencing depth for GRIT to produce full length transcript models, we merged a number of RNA-seq samples, e.g. all the samples from Larvae. These sample merges are specified in the complete GRIT configuration file used to execute the run, see Supplementary Data File 6. We note that GRIT, in its default mode thresholds alternative splicing events as follows: for each half-site (acceptor or donor site), reads crossing splice junctions are modeled only if the intron they cross is represented by at least 1% of the reads mapping to the half-site. To provide an example: if introns A and B share an acceptor sites, but have different donor sites, donor A and donor B respectively, then if the count of reads mapping to intron B is less than 1% of the count of reads mapping to intron A, intron B will not be modeled. Hence, alternative splicing events are only modeled if they are reasonably frequent in at least one sample. Our strategy is conservative: it is possible that we have not modeled rare or cell-type specific splicing events. This run resulted in 439,000 transcript models for 14,266 genes, including 72% of FB5.45 transcripts and 77% of FlyBase genes. These GRIT models also included gene merges at 1332 loci. We manually reviewed each gene merge to evaluate the cause. The majority of gene merges were due to incomplete 3' gene boundary information: missing polyadenylation sites resulted in 3' to 5' gene merges and hence long internal exons. This was not surprising, we have deep CAGE and RNA-seq data, but comparatively shallow 3' end gene boundary information: 1.84 M reads with poly(A) tails from poly(A) end enriched RNA-seq, and 32,000 3' ESTs and full-length cDNAs. After comprehensive manual review, we accepted 104 of the 1332 putative merges on the basis that these were mediated by uniquely mapping splice junction reads that passed filtration and were present in at least two biological replicates or samples. These analyses also lead us to look for gene merges between novel transcripts and known genes. We reviewed all gene models with known retained introns (5558 genes) and first exons that were longer than 5 kb or 1 kb longer than the longest FlyBase r.5.45 first exon at each gene (285 genes). We selected 71,015 transcripts for deletion and manually annotated an additional 207 novel genes that had unambiguous CAGE peaks (more than 10 reads in a primary peak), and more than 20x RNA-seq coverage across a putative gene-body, but no poly(A) read to provide 3' boundary information. In these manual cases, we selected the 3' boundary as the last base with

RNA-seq read coverage or, in higher coverage cases, the first base with a 100 fold drop in coverage.

To comprehensively identify regions with CAGE and RNA-seq data, but no poly(A) information, we ran a genome-wide scan for regions strong CAGE signal and proximal downstream RNA-seq signal. First, we trained a Random Forest (RF) (sklearn version 8.7.1) to identify 5' gene boundaries from CAGE peaks (a genomic position with the 5' ends of one or more CAGE tags aligned), RNA-seq, and genome sequence data using the Celniker full length cDNA collection (Supplementary Data File 4) as a positive training set, and CAGE peaks in CDS exons with no supporting EST or cDNA data as a negative training set (filtered CAGE tracks are given in Supplementary Data File 7), the covariates used to train our Random Forest Classifier are given in Supplementary Data File 8. The fitted classifier had sensitivity of 95% an FDR of 5% under cross validation on a held-out test set. However, we note that the purity of the negative control cannot be assured, and hence the true false discovery rate may be much lower. We fitted the classifier 100 times with randomly selected test sets to compute the variability of the imputed sensitivity and specificity, and found both to have standard deviations of 1%. We ran the RF genome-wide and classified all CAGE peaks as "candidate TSSs" or "Other". Next, we scanned all candidate TSSs for proximal RNA-seq signal, and subdivided regions into candidate single exon and multi-exonic genes. For candidate single exon genes, we required that no splice junction be present within 2 kb, that they have at least 20x mean coverage in our RNA-seq data and maximum coverage of at least 100x (over at least one nucleotide) within 2 kb of the CAGE peak, and the minimum RNA-seq coverage within the 2 kb region occur downstream of the maximum. For candidate spliced genes we looked for at least 20x mean RNA-seq coverage between the CAGE peak and a splice junction within 2 kb. These settings were based on extensive manual browsing and tuning. While we have attempted to be comprehensive, undoubtedly additional genes and transcripts remain to be discovered in our dataset. We note that our insistence on the presence of CAGE and RNA-seq data likely dramatically reduced the false discovery rate of the initial machine learning approach (described above) to peak-calling in CAGE data. This scan resulted in 7369 candidate single exon genes, all except 824 of which corresponded to annotated Transposable Elements (TEs) (overlapped an annotated element by >50%), and this filtered set (no TEs) we reviewed manually. We identified 1658 candidate spliced genes and reviewed each of these. These were not TE filtered prior to review on the basis that some TEs may be spliced into gene bodies, e.g. via recent exaptation (see below for additional TE filtering steps). This process resulted in the manual annotation of 3135 transcripts of 471 novel genes (678 manually annotated genes in total). As with GRIT, we built all possible transcript isoforms given our short read sequence data. We assigned gene transcript boundaries as the last contiguous base with RNA-seq coverage or after a 100 fold

fall-off in high expression cases. All GRIT and manual models for new genes were BLASTed against the FB5.45 Transposable Element sequence database, and all models with BLAST E-values < 0.0001 were removed from the annotation.

Finally we reviewed and recovered any missed known genes or transcripts in order to generate a comprehensive genome annotation. We reviewed our previous genome annotation efforts, which we now call modENCODE version 1 (MDv1) [3] and MDv2[14,15] to identify any gene or transcript models that were not reproduced in our GRIT and manual analysis. We also compared to FB5.45 and RefSeq (downloaded Feb. 2, 2013). This resulted in adding back a number of missed low expression genes as well as small RNA genes (e.g. tRNAs, miRNAs, etc.). These results are summarized in Supplementary Figure 1. The resulting complete annotation is MDv3 (Supplementary Data File 1), and includes attributions for each annotation.

6.6.11. Predicting proteins based on transcript models. In each transcript, we automatically annotated the longest ORF as a predicted protein whenever that ORF was at least 100 aa in length. When the longest ORF was between 20 aa and 100 aa, we evaluated each ORF longer than 20 aa as follows: we ran RPS-BLAST using the CDD (as below) and annotated any ORF with a CDD hit E-value of 1e-5 or less; we ran PhyloCSF (as below) and annotated any ORF with a conservation score of -0.2 or more. We note that this procedure identified novel conserved ORFs in 277 FB5.45 "non-coding" genes out of 893 such annotated genes, as well as 391 conserved ORFs in novel genes. In all, short conserved ORFs were identified in 27% of genes with no ORF over 100aa. Only 5% of these calls were due to the CDD RPS-BLAST search, the remainders were called by PhyloCSF. We consider these novel short ORFs "provisional"; extensive validation will be required to determine if they are translated in vivo.

6.6.12. siRNA analysis. The Drosophila melanogaster genome was segmented based on small RNA-seq read coverage of small RNA libraries in heads, ovaries, and testes. We clustered overlapping read regions into consensus segments and adjacent segments separated by less than 500 bp were then merged. The segments overlapping with TEs were excluded. cis-NAT siRNA features were extracted from these segments. Features used for the predictive model included 21 nt read frequency (21nt reads/all size reads); strand ratio (21 nt read ratio of sense/antisense); and read length distribution (mean, standard deviation, mode). We built a one-class predictive model which was trained on the previously published cisNAT siRNA loci from our and other labs[16,17,18,19] using the above features, and was applied to predict cis-NAT siRNAs on all segments genome-wide, separately for each library. In summary, minimum expression for annotating cis-NAT siRNA loci were 21nt reads $\geq$ 1 RPM for

both sense and antisense strands (5-95 percentile range is 2.9 – 70.4 RPM); minimum 21nt percentage (21nt reads/all size reads) for the calling siRNA loci was 60% (5-95 percentile range is 68% – 88%); minimum sense and antisense strand ratio was < 4.5 fold (5-95 percentile range is < 2.3 fold).

6.6.13. Conserved Domain and GO analysis of complex loci. We utilized the NCBI Conserve Domain Database (CDD)[20] and the Reverse Psi-BLAST (RPS-BLAST) tool[21] to identify functional domains in predicted proteins, using default settings. We used an E-value threshold of 1e-5 to specify potential hits. The Reverse Position Specific BLAST 2.2.26+ algorithm as part of the NCBI BLAST+ standalone package (version 2.2.26) was used to identify conserved domains within putative conserved domains.

To further characterize genes that express alternatively spliced transcripts, we examined conserved protein domains. Among genes with the capacity to produce more than 100 transcripts (292 genes), there are a number of significantly enriched conserved protein domains (FDR<1%), several corresponding to RNA binding domains: K homology, ELAV/HuD family splicing factor, sex-lethal family splicing factor, glycine-rich RNA-binding protein 4 motif, heterogeneous nuclear ribonucleoprotein R, Q family, and the half-pint family. A number of kinase-related domains are also strongly enriched. The most enriched Biological Process GO term is synaptic transmission (16 genes, FDR<7e-14).

6.6.13.1. *Identifying conserved ORFs that lack known domains.* We utilized the program PhyloCSF[22]to identify novel conserved ORFs that lacked known domains in the CDD database. The inputs to the algorithm are the 14 flies multiple alignment in MAF format (reviewed in[23])and the set of ORFs called by GRIT in our transcript annotation (see below, "Predicting proteins based on transcript models"). The algorithm was run in the "AsIs" mode which analyzes only the input ORFs (ORFs are not discovered by PhyloCSF). Based on communication with the Kellis group and their previous experience[24] (also, personal communication with Mike Lin), we utilized a conservation score threshold of -0.2 to identify conserved proteins.

6.6.14. Defining lncRNA elements. We defined lncRNA genes as those that lack any coding transcript given the above definition, and that encode no known small RNA (e.g. tRNAs, miRNAs, etc.). We note that this means that our annotation includes non-coding transcripts of coding genes. In Drosophila, there is one gene known to encode four 11aa ORFs[25] and hence it is possible that some of our lncRNAs may yet encode conserved and/or functional short polypeptides. However, PhyloCSF run time is exponential in minimum ORF length between 10 aa and 20 aa, due to an exponential increase in the number of

such ORFs present in transcript models. Furthermore, the power of the model is predicated on being able to observe protein-coding structure in multi-species alignments, e.g. third base wobble[22]This power is dampened in short ORFs, and after extensive manual review we determined that 20 aa was likely close to the limit of detection of the algorithm. This corresponds roughly to the limits of detection of MS/MS in our experience[26]and highlights the difficulty of identifying short protein coding sequencing, and the importance of emerging assays such as Ribo-seq[27]

6.6.15. MISO analysis of splicing dynamics. We parsed the annotation gtf file to generate GFF3 files containing individual splicing event annotations using a perl script described[14]. MISO[28]was used to quantitate the splicing events for all samples in single read mode as described in[14]

We identified 25,756 alternative splicing events in the transcript models. Of these, we focused on 17,447 events that produce only two isoforms per gene and do not have overlapping annotated features that might confound quantitation and analysis. We calculated Ψ values for each event in each tissue and developmental sample. We observed nearly identical distributions of median Ψ values for all events across all samples, among just the developmental samples and among just the tissue samples (Supplementary Fig. 4).

It has previously been shown that mammalian alternative exons whose magnitudes of splicing changes are large are more conserved and more frame-preserving than exons with low magnitude splicing changes[30]. To determine if this is also true in Drosophila, we characterized the conservation and reading-frame-preservation properties of cassette exons based on the magnitude of their tissue-specific regulation. We divided exons into three bins based on $\Delta\ \Psi$: high ($\Delta\ \Psi$>50%, n=395), moderate ($\Delta\ \Psi$ 25-50%, n=98) and low ($\Delta\ \Psi$<25%, n=68). Exons with high $\Delta\ \Psi$ are more conserved than those with moderate or low $\Delta\ \Psi$s, both within the exon and the flanking introns, in particular the upstream intron (Supplementary Fig. 11). In addition, we find that exons with high $\Delta\ \Psi$s tend to preserve the reading-frame more often than exons with moderate or low $\Delta\ \Psi$s (Supplementary Fig 10, chi-square p-value 1e-9, permutation test p-value 3e-8).

6.6.16. Detailed analysis of sex-specific splicing in somatic tissues. We previously identified hundreds of sex-specific splicing events from whole adult male and female RNA-seq data[6]. To further explore sex-specific splicing, we compared the splicing patterns in male and female heads. There were striking differences in gene expression levels between male and female heads, however, only six splicing events were consistently differentially spliced between males and females in heads at each time point after eclosion (average Δ

Ψ>20%). Of these, the strongest was the sex-specific 5' splice site in fruitless (avg. Δ Ψ=91%). Two other sex-specific splicing events occur in doublesex. The final three events were a retained intron in CG6236, an alternative 5' splice site in Ca^{2+}-channel protein α_1 subunit T and an alternative first exon in Septin 4. Of the other known splicing events in the sex-determination pathway, the 5' splice site in transformer had an average Δ Ψ of 48% (though one comparison had a Δ Ψ=16%), sex-specific splicing of male specific lethal-2 was not observed between male and female heads (avg. Δ Ψ=5%), and splicing events from Sex lethal (Sxl) were not quantified due to annotation complexity. When we conducted the quantification on a simplified set of transcripts (MDv1[3]), Sxl is the most sex-specific splicing event in the genome. Surprisingly, these results show that there is little sex-specific splicing in Drosophila heads.

6.6.16.1. *Differential Gene Expression Analysis.* Differential gene expression analysis was conducted only for our adult treatment samples. Our negative control used for this analysis the wild-type adult fly in gender-balanced mixed populations. Gene-level BPKMs were computed on independent biological replicates. We conducted quantile normalization of the BPKMs across all treatments and the negative control.

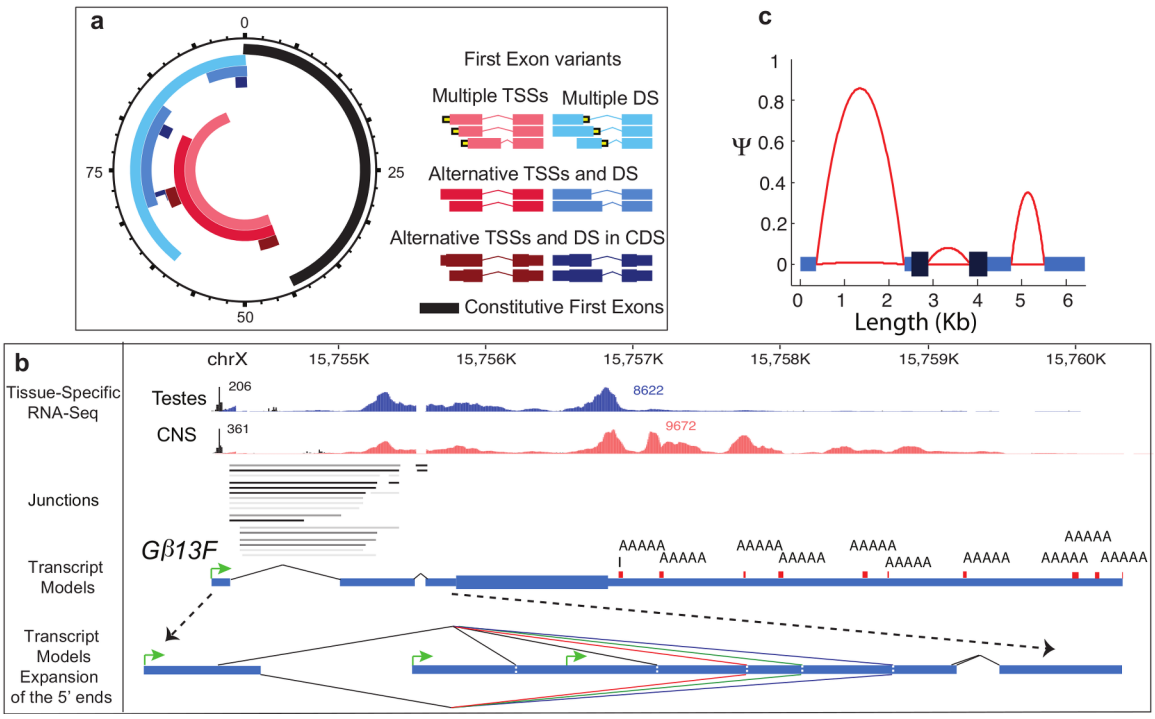

FIGURE 6.5.2. Splicing complexity across the gene body. a, Alternative first exons occur in two main configurations: multiple transcription start sites (TSS, pink) and multiple donor sites (DS, light blue). A subset of the genes in the multiple TSS category produce transcripts with different TSSs and shared DSs (red), and a subset of the genes in the DS category produce transcripts with a shared TSS and different DSs (blue). Some genes in the multiple TSS category directly affect the encoded protein (maroon), and similarly for DS (dark blue). Overlap of configurations is radially proportional (units indicate percentage of all spliced genes). b, Poly(A)+ testes (blue) and CNS (orange) stranded RNA-seq of $G\beta13F$ showing complex processing and splicing of the 5'UTR. Splice junctions (shaded gray as a function of usage) and an expansion of the 5'UTR showing some of the complexity. Transcription of the gene initiates from one of three different promoters (green arrows) terminates at one of ten possible polyA+ addition sites (from adult head poly(A)+seq, red) and generates 235 transcripts. The first exon has two alternative splice acceptors that splice to one of eleven different donor sites. Only five donor sites are shown due to the proximity of splice sites. Four splice donors are represented by the single red line differing by 12, 5 and 19bp respectively. Three splice donors are represented by the single green line differing by 12 and 11bp. Two splice donors are represented by the single purple line differing by 7bp. These splice variants are combined with four proximal internal splices (Supplementary Fig. 3a) to generate the full complement of transcripts. c, Intron retention rates (Ψ) across the gene body. The genome-wide mean lengths of exons and introns are connected by red parabolic arcs, which illustrate the upper and lower quartiles of intron retention (across all samples) for introns retained at or above 20 Ψ in at least one sample.

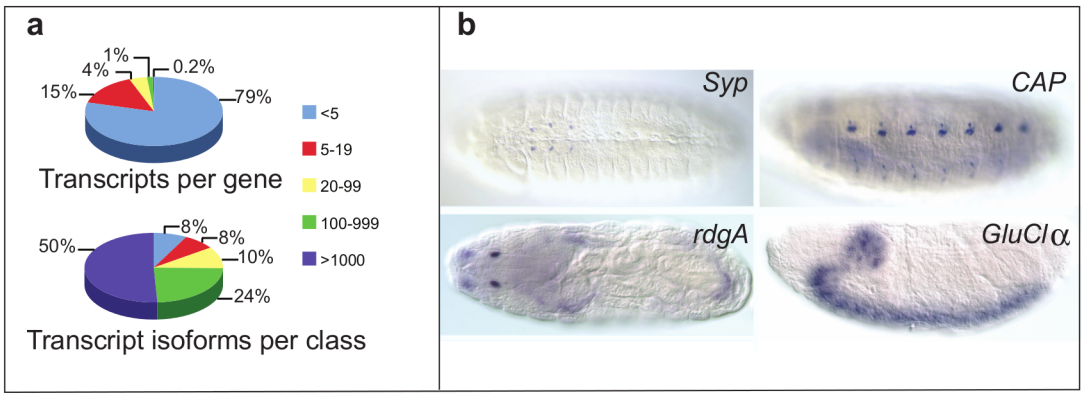

FIGURE 6.5.3. Complex splicing patterns are largely limited to neural tissues a, A small minority of genes (47, 0.2%) encode the majority of transcripts. b, In situ RNA staining of constitutive exons of four genes with highly complex splicing patterns in the embryo. Syncrip (Syp), Cap, Retinal degeneration A (rdgA) and GluClalpha show specific late embryonic neural expression in the ventral midline neurons; dorsal/lateral and ventral sensory complexes; Bolwig's organ or larval eye; and central nervous system respectively.

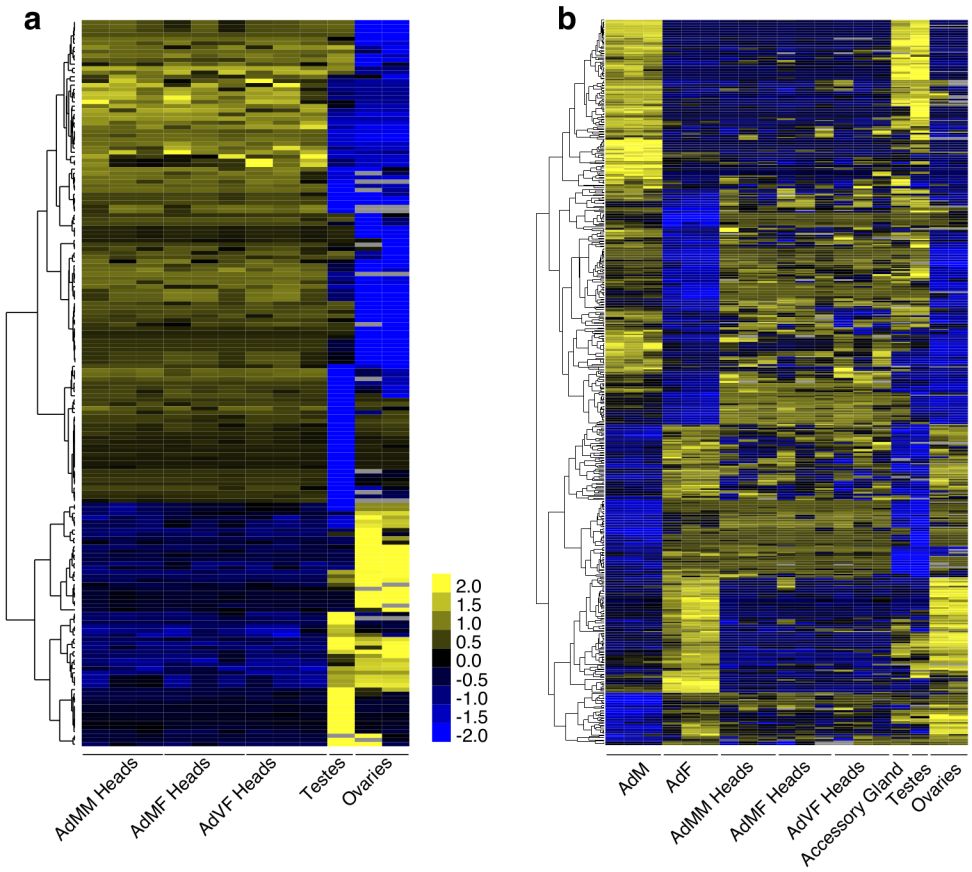

FIGURE 6.5.4. Sex-specific splicing is largely tissue-specific splicing a, Clusters of tissue-specific splicing events. The scale bar indicates Z-scores of Ψ. b, Sex-specific splicing events in whole animals are primarily testes- or ovary-specific splicing events.

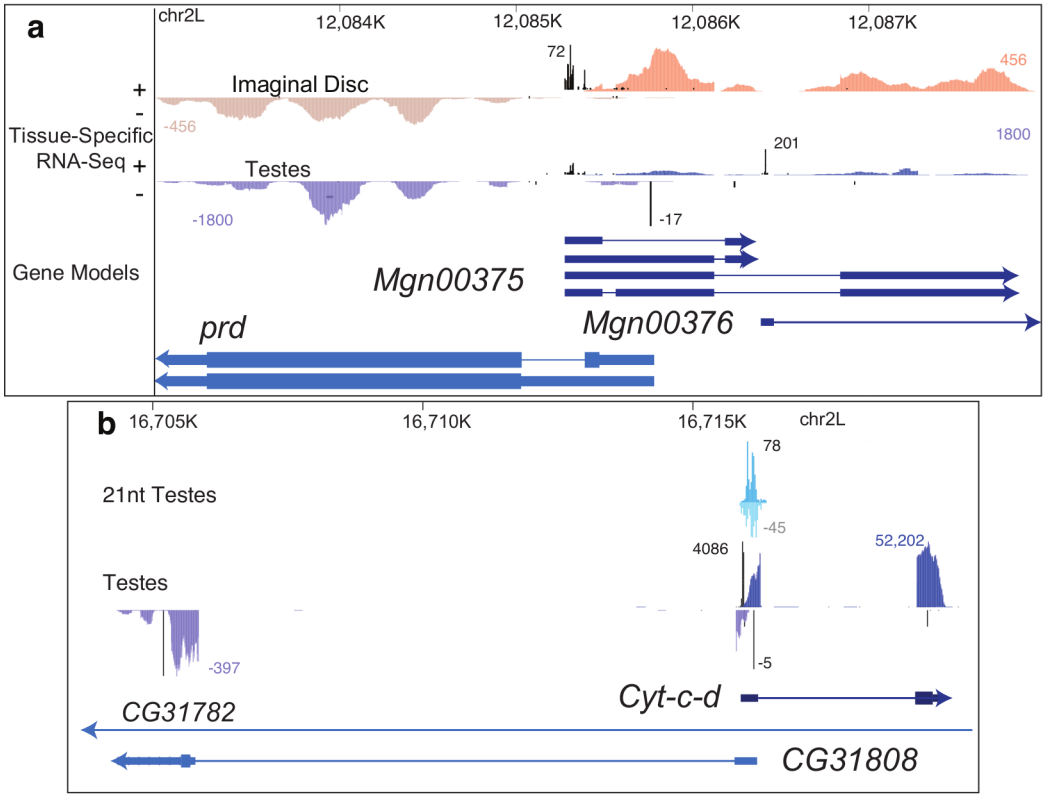

FIGURE 6.5.5. Examples of antisense transcription a, 5'/5' bidirectional antisense transcription at the prd locus. Short RNA sequencing does not reveal substantial siRNA (i.e. 21 nt-dominant small RNA) signal in this region (data not shown). b, A 5'/5' antisense region that produces substantial small RNA signal on both strands.

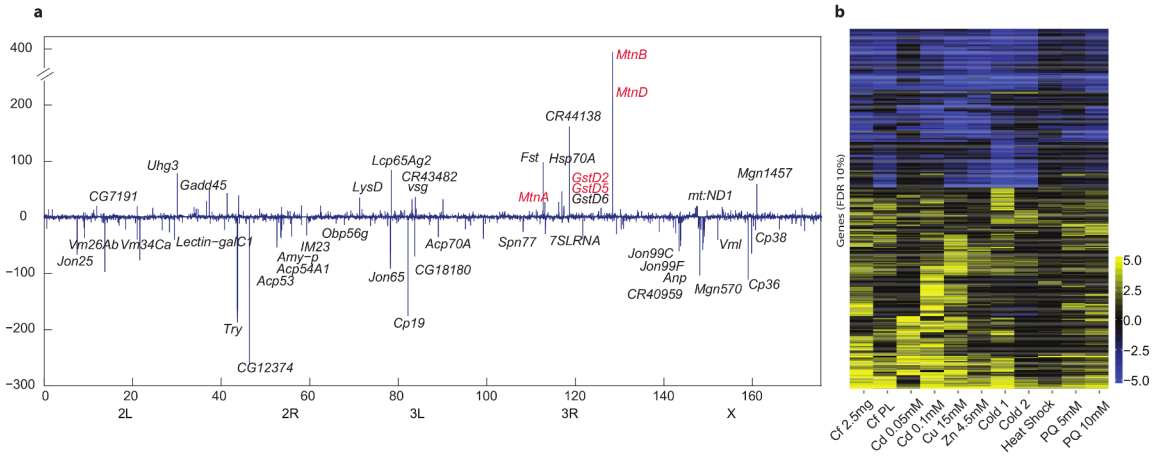

FIGURE 6.5.6. Effects of environmental perturbations on the Drosophila transcriptome. Adults were treated with caffeine (Cf), Cd, Cu, Zn, cold, heat, and paraquat (PQ). a, A genome-wide map of genes that are up or down regulated as a function of Cd treatment. Labeled genes are those that showed a 20-fold (<10% FDR) change in response (linear scale). Genes highlighted in red are those identified in larvae50. Some genes are omitted for readability, the complete figure and list of omitted genes are given in Supplementary Fig. 8a. b, Heat map showing the fold change of genes with an FDR<10% (differential expression) in at least one sample (log2 scale).

CHAPTER 7

Conclusion

We have developed two tools for the analysis of short read sequencing data. The first, statmap, is the first mapping tool that identifies candidate sequence under a probability model and is able to proive confidence bounds for mapping uncertainty. The second, GRIT, integrates multiple RNA data types and is able to identify and quantify novel genes, exons, introns, transcript bounds and, in some cases, transcripts. We hope that our models and tools will be useful for the biological community.

Bibliography

[1] Anders, S. and Huber, W. (2012). Differential expression of rna-seq data at the gene level–the deseq package. *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*.

[2] Batut, P. and Gingeras, T. R. (2013). Rampage: Promoter activity profiling by paired-end sequencing of 5'-complete cdnas. *Current Protocols in Molecular Biology*, pages 25B–11.

[3] Bickel, P. J. and Doksum, K. A. (2001). Mathematical statistics, volume i.

[4] Brown, J. B., Boley, N., Eisman, R., Mayd, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Park, S., Suzukii, A. M., Wan, K. H., Yub, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., Davis, C. A., Frise, E., Hammonds, A. S., Olson, S., Shenker, S., Sturgill, D., Andrews, J., Weng, J., Robinson, G., Hernandez, J., Bickel, P. J., Carninci, P., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Lai, E. C., Oliver, B., Graveley, B. R., and Celniker, S. E. (Submission ID 2012-12-15978A). Diversity and dynamics of the drosophila transcriptome. *Nature*.

[5] Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):94.

[6] Butler, J. E. and Kadonaga, J. T. (2002). The rna polymerase ii core promoter: a key component in the regulation of gene expression. *Genes & development*, 16(20):2583–2592.

[7] Celotto, A. M. and Graveley, B. R. (2001). Alternative splicing of the drosophila dscam pre-mrna is both temporally and spatially regulated. *Genetics*, 159(2):599–608.

[8] Cenik, C., Chua, H. N., Zhang, H., Tarnawsky, S. P., Akef, A., Derti, A., Tasan, M., Moore, M. J., Palazzo, A. F., and Roth, F. P. (2011). Genome analysis reveals interplay between 5' utr introns and nuclear mrna export for secretory and mitochondrial genes. *PLoS genetics*, 7(4):e1001366.

[9] Collins, J. E., White, S., Searle, S. M., and Stemple, D. L. (2012). Incorporating rna-seq data into the zebrafish ensembl genebuild. *Genome research*, 22(10):2067–2078.

[10] Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., et al. (2010). A comparison between ribo-minus rna-sequencing and polya-selected rna-sequencing. *Genomics*, 96(5):259–265.

[11] Dahl, J. and Vandenberghe, L. (2006). Cvxopt: A python package for convex optimization. In *Proc. Eur. Conf. Op. Res.*

[12] Delhomme, N., Padioleau, I., Furlong, E. E., and Steinmetz, L. M. (2012). easyrnaseq: a bioconductor package for processing rna-seq data. *Bioinformatics*, 28(19):2532–2533.

[13] Di Ruscio, A., Ebralidze, A. K., Benoukraf, T., Amabile, G., Goff, L. A., Terragni, J., Figueroa, M. E., Pontes, L. L. D. F., Alberich-Jorda, M., Zhang, P., et al. (2013). Dnmt1-interacting rnas block gene-specific dna methylation. *Nature*.

[14] Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM.

[15] Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016.

[16] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. *Nucleic acids research*, page gkt1196.

[17] Gelbart, W., Crosby, M., Matthews, B., Chillemi, J., Russo Twombly, S., Emmert, D., Bayraktaroglu, L., Smutniak, F., Kossida, S., Ashburner, M., et al. (1999). The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Research*, 27.

[18] Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L., Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H., Auerbach, R. K., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G., Good, P. J., Green, P., Guigo, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang, H., Hubbard, T. J. P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H., Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R., and Waterston, R. (Submission ID 2012-12-15985A). Comparison of 3 metazoan transcriptomes. *Nature*.

[19] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652.

[20] Grant, M., Boyd, S., and Ye, Y. (2008). Cvx: Matlab software for disciplined convex programming.

[21] Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., et al. (2010). The developmental transcriptome of drosophila melanogaster. *Nature*, 471(7339):473–479.

[22] Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, 28(5):503–510.

[23] Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12):e131–e131.

[24] Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: The reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

[25] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

[26] Hilgers, V., Perry, M. W., Hendrix, D., Stark, A., Levine, M., and Haley, B. (2011). Neural-specific elongation of 3' utrs during drosophila development. *Proceedings of the National Academy of Sciences*, 108(38):15864–15869.

[27] Hillier, L. W., Reinke, V., Green, P., Hirst, M., Marra, M. A., and Waterston, R. H. (2009). Massively parallel sequencing of the polyadenylated transcriptome of c. elegans. *Genome research*, 19(4):657–666.

[28] Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. (2007). Wolf psort: protein localization predictor. *Nucleic acids research*, 35(suppl 2):W585–W587.

[29] Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., Yu, C., Booth, B. W., Zhang, D., Wan, K. H., et al. (2011). Genome-wide analysis of promoter architecture in drosophila melanogaster. *Genome research*, 21(2):182–192.

[30] Jenkins, C., Michael, D., Mahendroo, M., and Simpson, E. (1993). Exon-specific northern analysis and rapid amplification of cdna ends (race) reveal that the proximal promoter ii (pii) is responsible for aromatase cytochrome p450 cyp19 expression in human ovary. *Molecular and cellular endocrinology*, 97(1):R1–R6.

[31] Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in rna-seq. *Bioinformatics*, 25(8):1026–1032.

[32] Juneau, K., Nislow, C., and Davis, R. W. (2009). Alternative splicing of ptc7 in saccharomyces cerevisiae determines protein localization. *Genetics*, 183(1):185–194.

[33] Kawakami, A., Kataoka, H., Oka, T., Mizoguchi, A., Kimura-Kawakami, M., Adachi, T., Iwami, M., Nagasawa, H., Suzuki, A., and Ishizaki, H. (1990). Molecular cloning of the bombyx mori prothoracicotropic hormone. *Science*, 247(4948):1333–1335.

[34] Lawless, C., Pearson, R. D., Selley, J. N., Smirnova, J. B., Grant, C. M., Ashe, M. P., Pavitt, G. D., and Hubbard, S. J. (2009). Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC genomics*, 10(1):7.

[35] Li, B. and Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

[36] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500.

[37] Lifton, R., Goldberg, M., Karp, R., and Hogness, D. (1978). The organization of the histone genes in drosophila melanogaster: functional and evolutionary implications. In *Cold Spring Harbor symposia on quantitative biology*, volume 42, pages 1047–1051. Cold Spring Harbor Laboratory Press.

[38] Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S. D., Mis, E., Zegar, C., Gutwein, M. R., Khivansara, V., et al. (2010). The landscape of c. elegans 3' utrs. *Science*, 329(5990):432–435.

[39] Marygold, S. J., Leyland, P. C., Seal, R. L., Goodman, J. L., Thurmond, J., Strelets, V. B., Wilson, R. J., et al. (2013). Flybase: improvements to the bibliography. *Nucleic acids research*, 41(D1):D751–D757.

[40] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

[41] Penalva, L. O. and Sánchez, L. (2003). Rna binding protein sex-lethal (sxl) and control of drosophila sex determination and dosage compensation. *Microbiology and molecular biology reviews*, 67(3):343–359.

[42] Pilanci, M., El Ghaoui, L., and Chandrasekaran, V. (2012). Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems*, volume 25, pages 2429–2437.

[43] Rewitz, K. F., Yamanaka, N., Gilbert, L. I., and O'Connor, M. B. (2009). The insect neuropeptide ptth activates receptor tyrosine kinase torso to initiate metamorphosis. *Science*, 326(5958):1403–1405.

[44] Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12(1):480.

[45] Roberts, A. and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1):71–73.

[46] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., et al. (2010). De novo assembly and analysis of rna-seq data. *Nature methods*, 7(11):909–912.

[47] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[48] Rojas-Duran, M. F. and Gilbert, W. V. (2012). Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA*, 18(12):2299–2305.

[49] Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., and Zipursky, S. L. (2000). Drosophila dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6):671–684.

[50] Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092.

[51] Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009–1014.

[52] Smibert, P., Miura, P., Westholm, J. O., Shenker, S., May, G., Duff, M. O., Zhang, D., Eads, B. D., Carlson, J., Brown, J. B., et al. (2012). Global patterns of tissue-specific alternative polyadenylation in< i> drosophila</i>. *Cell reports*, 1(3):277–289.

[53] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515.

[54] Tu, Q., Cameron, R. A., Worley, K. C., Gibbs, R. A., and Davidson, E. H. (2012). Gene structure in the sea urchin strongylocentrotus purpuratus based on transcriptome analysis. *Genome Research*, 22(10):2079–2087.

[55] Uhlmann, J. K. (1991). Satisfying general proximity/similarity queries with metric trees. *Information processing letters*, 40(4):175–179.

[56] Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investments: Maximum likelihood solutions for positive linear inverse problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 569–612.

[57] Wang, Z., Gerstein, M., and Snyder, M. (2009a). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[58] Wang, Z., Gerstein, M., and Snyder, M. (2009b). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63.

[59] Wojtowicz, W. M., Flanagan, J. J., Millard, S. S., Zipursky, S. L., and Clemens, J. C. (2004). Alternative splicing of< i> drosophila</i> dscam generates axon guidance receptors that exhibit isoform-specific homophilic binding. *Cell*, 118(5):619–633.

[60] Yook, K., Harris, T. W., Bieri, T., Cabunoc, A., Chan, J., Chen, W. J., Davis, P., De La Cruz, N., Duong, A., Fang, R., et al. (2012). Wormbase 2012: more genomes, more data, new website. *Nucleic acids research*, 40(D1):D735–D741.

[61] Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, 12(1):290.

1 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628, (2008).

2 Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344-1349, (2008).

3 Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol 786, 181-200, doi:10.1007/978-1-61779-292-2_11 (2012).

4 Mangone, M. et al. The landscape of C. elegans 3'UTRs. Science 329, 432-435, (2010).

5 Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 469, 97-101, (2011).

6 Graveley, B. R. et al. The developmental transcriptome of Drosophila melanogaster. Nature 471, 473-479, (2011).

7 Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7, 562-578, (2012).

8 Collins, J. E., White, S., Searle, S. M. & Stemple, D. L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. Genome Res 22, 2067-2078, (2012).

9 Carninci, P. et al. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res 13, 1273-1289, (2003).

10 Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22, 1760-1774, (2012).

11 Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101-108, (2012).

12 Hoskins, R. A. et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21, 182-192, (2011).

13 Cherbas, L. The Transcriptional Diversity of 25 Drosophila Cell Lines. Genome Res (2010).

14 Boley, N. et al. Genome guided transcript construction from integrative analysis of RNA sequence data. Nature Biotechnology (2013).

15 Celniker, S. E. & Rubin, G. M. The Drosophila melanogaster genome. Annu Rev Genomics Hum Genet 4, 89-117 (2003).

16 Stapleton, M. et al. The Drosophila Gene Collection: Identification of Putative Full-Length cDNAs for 70% of D. melanogaster Genes. Genome Res 12, 1294-1300 (2002).

17 Spradling, A. C. et al. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics 153, 135-177 (1999).

18 Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476, (2008).

19 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40, 1413-1415, (2008).

20 Smibert, P. et al. Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep 1, 277-289, doi:10.1016/j.celrep.2012.01.001 (2012).

21 St Laurent, G. et al. Genome-wide analysis of A-to-I RNA editing by single-molecule sequencing in Drosophila. Nat Struct Mol Biol, (2013).

22 Telonis-Scott, M., Kopp, A., Wayne, M. L., Nuzhdin, S. V. & McIntyre, L. M. Sex-specific splicing in Drosophila: widespread occurrence, tissue specificity and evolutionary conservation. Genetics 181, 421-434, (2009).

23 Hartmann, B. et al. Distinct regulatory programs establish widespread sex-specific alternative splicing in Drosophila melanogaster. RNA 17, 453-468, (2011).

24 Chang, P. L., Dunham, J. P., Nuzhdin, S. V. & Arbeitman, M. N. Somatic sex-specific transcriptome differences in Drosophila revealed by whole transcriptome sequencing. BMC Genomics 12, 364, (2011).

25 Matthews, L. et al. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37, D619-622, (2009).

26 Lipshitz, H. D., Peattie, D. A. & Hogness, D. S. Novel transcripts from the Ultrabithorax domain of the Bithorax Complex. Genes and Development 1, 307-322 (1987).

27 Tupy, J. L. et al. Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster. Proc Natl Acad Sci U S A 102, 5495-5500 (2005).

28 Young, R. S. et al. Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome Biol Evol 4, 427-442, (2012).

29 Kondo, T. et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9, 660-665, (2007).

30 Katayama, S. et al. Antisense transcription in the mammalian transcriptome. Science 309, 1564-1566, (2005).

31 Derrien, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22, 1775-1789, (2012).

32 Duncan, D. M., Burgess, E. A. & Duncan, I. Control of distal antennal identity and tarsal development in Drosophila by spineless-aristapedia, a homolog of the mammalian dioxin receptor. Genes Dev 12, 1290-1303 (1998).

33 Schwartz, C., Locke, J., Nishida, C. & Kornberg, T. B. Analysis of cubitus interruptus regulation in Drosophila embryos and imaginal disks. Development 121, 1625-1635 (1995).

34 Misra, S. et al. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biology 3, research0083 (2002).

35 Lipovich, L. et al. Activity-dependent human brain coding/noncoding gene regulatory networks. Genetics 192, 1133-1148, (2012).

36 Okamura, K. & Lai, E. C. Endogenous small interfering RNAs in animals. Nat Rev Mol Cell Biol 9, 673-678, (2008).

37 Okamura, K., Balla, S., Martin, R., Liu, N. & Lai, E. C. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nat Struct Mol Biol 15, 581-590, (2008).

38 Czech, B. et al. An endogenous small interfering RNA pathway in Drosophila. Nature 453, 798-802, (2008).

39 Ghildiyal, M. et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 320, 1077-1081, (2008).

40 Engstrom, P. G. et al. Complex Loci in human and mouse genomes. PLoS Genet 2, e47, doi:10.1371/journal.pgen.0020047 (2006).

41 Whitehead, A. et al. Genomic and physiological footprint of the Deepwater Horizon oil spill on resident marsh fishes. Proc Natl Acad Sci U S A 109, 20298-20302, (2012).

42 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218-223, (2009).

43 Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 22, 1646-1657, (2012).

44 Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154, 240-251, (2013).

45 Kosti, I., Radivojac, P. & Mandel-Gutfreund, Y. An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors. PLoS Comput Biol 8, (2012).

46 Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7, 542-561, (2012).

47 Bickel, P. J., Boley, N., Brown, J. B., Huang, H. & Zhang, N. R. Subsampling methods for genomic inference. Ann. Appl. Stat. 4, 1660-1697 (2010).

48 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7, 1009-1015, (2010).

49 Anders, S. & Huber, W. Differential expression analysis for sequence count data. Genome Biol 11, R106, (2010).

50 Yepiskoposyan, H. et al. Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. Nucleic Acids Res 34, 4866-4877, (2006).

S1 Celniker, S.E., et al. Finishing a whole-genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biology. 3(12)1-14. (2002)

S2 Maroni G, Stamey SC. Use of blue food to select synchronous late third-instar larvae. Drosophila Inf Serv. 59:142-143. (1983)

S3 Graveley, B. R. et al. The developmental transcriptome of Drosophila melanogaster. Nature 471, 473-479, doi:nature09715 [pii]

S4 Functional Implications of the Unusual Spatial Distribution of a Minor a-Tubulin lsotype in Drosophila: A Common Thread among Chordotonal Ligaments, Developing Muscle, and Testis Cyst Cells Kathleen A. Matthews, David F. B. Miller, and Thomas C. Kaufman. Developmental Biology 1990, 137, 171-183.

S5 Fristrom JW, Mitchell HK. The preparative isolation of imaginal discs from larvae of Drosophila melanogaster. J Cell Biol. 27(2):445-8. (1965)

S6 Smibert, P. et al. Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep 1, 277-289, doi:10.1016/j.celrep.2012.01.001 (2012).

S7 Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101-108, doi:nature11233

S8 Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc 7, 542-561, doi:nprot.2012.005 [pii]

S9 Hoskins, R. A. et al. Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21, 182-192, doi:gr.112466.110 [pii]

S10 Wan, K. H. et al. High-throughput plasmid cDNA library screening. Nat Protoc 1, 624-632 (2006).

S11 Boley, N., et al. Ab initio transcript assembly using gene boundary data. Submitted to Nature Biotech. (2013)

S12 Retelska, D., Iseli, C., Bucher, P., Jongeneel, C. V. & Naef, F. Similarities and differences of polyadenylation signals in human and fly. BMC Genomics 7, 176, doi:1471-2164-7-176 [pii]

S13 Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37, W202-208, doi:gkp335 [pii]

S14 Gerstein, M. et al. Comparison of the transcriptomes of flies, human and worms. Nature. Submission ID: 2012-12-15978A. (2013).

S15 Oliver, B. Comparative trancriptome analysis using 20 Fly Species. Manuscript in preparation. (2013).

S16 Okamura, K. & Lai, E.C. Endogenous small interfering RNAs in animals. Nat Rev Mol Cell Biol 9, 673-8 (2008).

S17 Okamura, K., Balla, S., Martin, R., Liu, N. & Lai, E.C. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila. Nat Struct Mol Biol 15, 581-590 (2008).

S18 Czech, B. et al. An endogenous siRNA pathway in Drosophila. Nature 453, 798-802 (2008).

S19 Ghildiyal, M. et al. Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. Science 320, 1077-1081 (2008).

S20 Marchler-Bauer, A. et al. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res 41, D348-352, doi:gks1243 [pii]

S21 Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32, W327-331, doi:10.1093/nar/gkh454

S22 Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 27, i275-282, doi:btr209 [pii]

S23 Blankenberg, D., Taylor, J. & Nekrutenko, A. Making whole genome multiple alignments usable for biologists. Bioinformatics 27, 2426-2428, doi:btr398 [pii]

S24 Stark, A. et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450, 219-232, doi:nature06340 [pii]

S25 Kondo, T. et al. Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9, 660-665, doi:ncb1595 [pii]

S26 Banfai, B. et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 22, 1646-1657, doi:22/9/1646 [pii]

S27 Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218-223, doi:1168978

S28 Katz Y., Wang E.T., Airoldi E.M., Burge C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation.Nat Methods. 7(12):1009-15. doi: 10.1038/nmeth.1528. (2010)

S29 Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 15545–15550. (doi:10.1073/pnas.0506580102) (2005)

S30 Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476, (2008).