Optimization of RNA-Sequencing Analysis and a Role for the Epidermis in Sensation

Claire Rankin Williams

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jay Z. Parrish, Chair

David Raible

Cole Trapnell

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Optimization of RNA-Sequencing Analysis and a Role for the Epidermis in Sensation

Claire Rankin Williams

Chair of the Supervisory Committee:

Associate Professor Jay Z. Parrish

Department of Biology

Our skin is our largest organ, covering our entire bodies, and is the first point of contact for numerous external stimuli. Recent evidence in mammalian models suggests that epidermal cells that comprise our skin may be playing an active role in sensing and responding to these stimuli, and for this dissertation, I sought to understand how widely conserved this epidermal sensation is. To begin, I used transcriptome analysis of different cell types in the skin to identify how skin may be able to respond to different stimuli. However, RNA-Sequencing (RNA-Seq) is still a relatively new technique with disparate analysis pipelines in place, and so I first set out to optimize methods for RNA-Seq analysis, which I did over the course of multiple bioinformatics projects. I assayed the effects of varying parameters for quality-based trimming on gene expression estimates, identifying that short reads remaining after trimming are the main driver of bias in expression estimates. I quantified the performance of 495 different differential expression workflows, using the metrics of recall and precision to describe a performance tradeoff observed across workflows. With current sequencing costs still a limiting factor for many bulk RNA-Seq experiments, I compared the tradeoff of sample size versus read depth for 30 high performance differential expression workflows and found inflection points at which performance degraded substantially. Together, these three projects allow me to make three broad recommendations. First, aggressive quality-based trimming is detrimental to RNA-Seq expression estimates and therefore should be avoided.

Second, the choice of analysis tools at each step of differential expression analysis will affect the results obtained and therefore tools chosen should be tailored to the specific question being asked. Finally, differential expression workflows perform better with larger sample sizes and higher read depths but when the two are constrained, obtaining sample sizes of at least six per group should be prioritized. With these methods in place, I contributed to RNA-Seq analyses for multiple collaborators, which are presented in this work in short vignettes. After describing these transcriptome methods optimizations and applications, I return to the initial topic of epidermal sensation. By using an invertebrate, *Drosophila melanogaster*, I show that the ability of epidermal cells to induce aversive behaviors and respond directly to mechanical stimuli is conserved across evolution. In fruit flies and in mammals, epidermal cells ensheathe the dendrites of somatosensory neurons and these epidermally-induced behaviors are mediated by canonical neuronal circuitry. Overall, I describe a novel function for invertebrate epidermal cells and show that *D. melanogaster* provides a fruitful system for investigating epidermal-neuronal interactions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I thank the many people who supported me and contributed to this work throughout my graduate research. In particular, I am grateful to my mentor, Dr. Jay Z. Parrish, for his enthusiastic support and encouragement of my PhD research and scientific growth. I thank my thesis committee members, Dr. Charlie Kim, Dr. David Raible, Dr. Cole Trapnell, Dr. Celeste Berg, and Dr. Rachel Wong for providing guidance and advice during the course of my studies. I had the opportunity to work with many wonderful collaborators and I would like to specifically thank those with whom I worked most closely: Dr. Alyssa Baccarella, Dr. Jiro Yoshino, Dr. Kazuo Emoto, and Dr. Peter Soba. I thank my colleagues in the Parrish lab for making the lab an exciting and stimulating place to do science. Finally, I thank my friends and family who supported me outside the lab so that I could be more effective during my time in lab.

Chapter 1 : INTRODUCTION

Neurobiologists have long been fascinated by neurons, the excitable cells that we use to sense, comprehend, and respond to the numerous stimuli we interact with daily. Yet neurons do not exist in isolation. They are surrounded by glial cells, immune cells, muscles, and epidermal cells, all working together to shape our interactions with the outside world. For my thesis research, I sought to explore contributions from these non-neuronal cell types to somatosensation and began by optimizing an emerging technique, RNA-Sequencing (RNA-Seq) for comprehensive profiling of different cell types. In this chapter, I first briefly introduce this key technique and second succinctly provide background on non-neuronal cells of the skin.

1.1 RNA-Seq: A rapidly emerging technique

RNA-Seq is a rapidly emerging technique used to obtain a comprehensive, quantitative description of all the genes or transcripts expressed in a given cell, cell type, tissue, or organism. The scale of information available from this technique is tremendous, with a single next generation sequencing (NGS) run often generating upwards of 300 million short reads that then must be processed to create expression estimates for every transcript or gene. RNA-Seq has replaced the previous standard in the field of molecular biology, microarray analysis, at an astounding rate over the course of my dissertation research: a PubMed search for the term “rna-seq” or “rnaseq” shows approximately 2100 results for all papers published through 2012 and almost 8000 results for 2018 alone [1]. However, with this widespread adoption has come great variability in the analysis methods and tools used by different labs and even different scientists within the same research group [2–4]. This variability can create confusion and difficulty in interpreting results when comparing across studies, especially since justification for the tools and techniques chosen is rare. As new analysis tools continue to be introduced - there are currently over 100 frequently used RNA-Seq analysis tools - it will be important for the field to develop consensus for the optimal way(s) to analyze RNA-Seq data.

In chapters two through four of this thesis, I describe three extensive studies I completed to provide guidance on how to analyze RNA-Seq data. As more scientists begin to use transcriptomic

analyses in their research, there is a critical need for bioinformaticians with the biological knowledge base and computational skills to interpret these big datasets. In the fifth chapter, I briefly describe three collaborations I undertook in this capacity, applying my RNA-Seq analysis expertise to various biological questions. Additionally, my proficiency in RNA-Seq analysis provided the foundation for me to launch my own project exploring the contributions of various non-neuronal cell types, particularly the epidermis, to sensation; these studies are presented in the sixth chapter.

1.2 Skin: An underappreciated sensory structure

Our skin is our largest organ, coating our entire bodies and providing our first point of contact with any external stimuli. Mammalian skin, composed of a multilayer stratified epidermis, specialized immune cells, and ramified somatosensory neuron processes, is known to serve many roles in daily life. Skin acts as a barrier, separating inside the body from outside and preventing the passage of external harmful agents or the evaporation of internal hydration [5]. Skin functions as a guardian, surveilling the periphery for viruses or other infectious microbes that may have penetrated the epidermal barrier, allowing for the swift mounting of an immune response [6]. Skin plays a role as a scaffold, providing a substrate over and through which neurons can elaborate dendrites [7]. Recent evidence suggests that the skin also contributes directly to somatosensation.

Sensory functions are ascribed to multiple non-neuronal peripheral cell types in the mammalian skin. Keratinocytes, the predominant epidermal cells in the skin at about 95% of all cells, depolarize in response to mechanical stimuli [8,9]. This depolarization may be functionally relevant as optogenetic stimulation of mice expressing Channelrhodopsin in keratinocytes can induce avoidance behaviors, albeit at low frequency [9]. Merkel cells, a specialized epidermal cell found in structures called touch domes located in regions of high tactile acuity, depolarize in response to mechanical stimuli and release the neurotransmitter norepinephrine at a synapse-like structure onto sensory afferent neurons [10]. Outside of epidermal cells, peripheral glia ensheathe somatosensory neurons and potentiate nociceptive behaviors when depolarized [11]. These findings suggest that we as neuroscientists should be looking beyond neurons to understand how stimuli are sensed, comprehended, and responded to. To our knowledge, no such sensory functions have been described for non-neuronal peripheral cells in an invertebrate, raising the question of how broadly evolutionarily conserved this role is. In the final chapter

of this thesis, I describe our efforts to address this question using the invertebrate *Drosophila melanogaster*.

1.3 References

1. pubmeddev. Home - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/.

2. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE* **12**, e0190152 (2017).

3. Ballouz, S., Dobin, A., Gingeras, T. R. & Gillis, J. The fractured landscape of RNA-seq alignment: the default in our STARs. *Nucleic Acids Res.* **46**, 5125–5138 (2018).

4. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, (2016).

5. Elias, P. M. Skin barrier function. *Curr. Allergy Asthma Rep.* **8**, 299–305 (2008).

6. Eyerich, K. & Eyerich, S. Immune response patterns in non-communicable inflammatory skin diseases. *J. Eur. Acad. Dermatol. Venereol*. **32**, 692–703 (2018).

7. Talagas, M., Lebonvallet, N., Leschiera, R., Marcorelles, P. & Misery, L. What about physical contacts between epidermal keratinocytes and sensory neurons? *Exp. Dermatol*. **27**, 9–13 (2018).

8. Fuchs, E. Keratins and the Skin. *Annu. Rev. Cell Dev. Biol*. **11**, 123–154 (1995).

9. Baumbauer, K. M. et al. Keratinocytes can modulate and directly initiate nociceptive responses. *eLife* **4**, (2015).

10. Hoffman, B. U. et al. Merkel Cells Activate Sensory Neural Pathways through Adrenergic Synapses. *Neuron* **100**, 1401–1413.e6 (2018).

11. Abdo, H. et al. Specialized cutaneous Schwann cells initiate pain sensation. *Science* **365**, 695–699 (2019).

Chapter 2 : TRIMMING OF SEQUENCE READS ALTERS RNA-SEQ GENE EXPRESSION ESTIMATES

Chapter 2 is adapted with minimal modification from:

Williams CR, Baccarella A, Parrish JZ, Kim CC. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 17(1): 103.

2.1 Abstract

2.1a Background

High-throughput RNA-Sequencing (RNA-Seq) has become the preferred technique for studying gene expression differences between biological samples and for discovering novel isoforms, though the techniques to analyze the resulting data are still immature. One pre-processing step that is widely but heterogeneously applied is trimming, in which low quality bases, identified by the probability that they are called incorrectly, are removed. However, the impact of trimming on subsequent alignment to a genome could influence downstream analyses including gene expression estimation; we hypothesized that this might occur in an inconsistent manner across different genes, resulting in differential bias.

2.1b Results

To assess the effects of trimming on gene expression, we generated RNA-Seq data sets from four samples of larval *Drosophila melanogaster* sensory neurons, and used three trimming algorithms—SolexaQA, Trimmomatic, and ConDeTri—to perform quality-based trimming across a wide range of stringencies. After aligning the reads to the *D. melanogaster* genome with TopHat2, we used Cuffdiff2 to compare the original, untrimmed gene expression estimates to those following trimming. With the most aggressive trimming parameters, over ten percent of genes had significant changes in their estimated expression levels. This trend was seen with two additional RNA-Seq data sets and with alternative differential expression analysis pipelines. We found that the majority of the expression changes could be mitigated by imposing a minimum length filter following trimming, suggesting that the differential gene

expression was primarily being driven by spurious mapping of short reads. Slight differences with the untrimmed data set remained after length filtering, which were associated with genes with low exon numbers and high GC content. Finally, an analysis of paired RNA-seq/microarray data sets suggests that no or modest trimming results in the most biologically accurate gene expression estimates.

2.1c Conclusions

We find that aggressive quality-based trimming has a large impact on the apparent makeup of RNA-Seq-based gene expression estimates, and that short reads can have a particularly strong impact. We conclude that implementation of trimming in RNA-Seq analysis workflows warrants caution, and if used, should be used in conjunction with a minimum read length filter to minimize the introduction of unpredictable changes in expression estimates.

2.2 Background

Within the past decade, RNA sequencing (RNA-Seq) has supplanted microarrays as the preferred technique for gene expression analysis. A typical workflow for RNA-Seq analysis involves aligning reads to an annotated genome followed by estimation of gene-level and/or isoform-level expression. In many cases, this is followed by statistical identification of genes that are differentially expressed between two or more sample groups. However, RNA-Seq presents unique analytical challenges, and accurate and robust tools to analyze sequencing data are rapidly evolving. As a result, analysis workflows can vary widely between studies.

One initial step of RNA-Seq analysis is to evaluate sequence read quality, which can vary substantially based on factors related to nucleic acid library preparation (*e.g.*, adapter contamination, polymerase errors) and sequencing (*e.g.*, cluster density, optical detection errors, phasing errors) [1]. For example, during library preparation, random hexamers are sometimes used as primers for double stranded cDNA synthesis, which leads to biases in nucleotide composition at the beginning of reads [2]. A second, intrinsic problem of sequencing by synthesis is phasing: different fragments within a cluster fall out of phase with one another as a result of slight differences in the timing of polymerization. Errors in phasing accumulate over time; thus, read quality tends to decrease toward the ends of sequence reads.

Further, errors have a tendency to co-occur, such that reads with two errors are more common than would be predicted based on a model in which errors occur independently of one another [3].

In the absence of pre-processing, phasing and other sequencing errors can lead to inclusion of incorrect base calls and, consequently, to erroneous read alignment. Current next generation sequencing technologies produce reads as short as 25 bases up to hundreds of bases; sequencing errors are less frequent in the shorter read data sets, but the proportional impact of a single incorrect base may be larger. Sequencing-associated errors are aggregated into a quality score that reflects the probability that a given base has been called incorrectly. Most common among these, the Phred quality score (Q) used in the Illumina platforms ranges from 0 to 40, with increasing scores corresponding to higher quality base calls; for example, a Q score of 40 represents a 1 in 10,000 chance that a base has been called incorrectly [4]. Similar quality scores are produced with alternative sequencing platforms as well. During pre-processing, the quality score can be used to eliminate poor quality bases that typically occur at the ends of reads, in a procedure commonly referred to as "trimming". This quality-based trimming is distinct from adapter trimming, which can be used to remove high quality internal bases matching the sequencing adapters used in library preparation [5]. Numerous approaches to quality-based trimming exist [6], all with the end result of retaining high quality internal bases while removing lower quality flanking bases.

However, as for pre-processing in general, quality-based trimming of reads is widely, but heterogeneously, applied. Thus, the specific algorithms and parameters used for quality score-driven trimming are a major determinant of what portions of reads are retained for further analysis. A broad survey of the major trimming algorithms currently in use found that although trimming prior to mapping of RNA-Seq reads leads to a decrease in the total number of reads, it concurrently increases the proportion of the remaining reads that map, suggesting that trimming is effective in removing reads that could not be mapped to the reference genome [6].

Although the above study suggested that trimming is beneficial, multiple lines of evidence suggest that it can also have detrimental effects. First, while errors in the assembly of a known transcriptome decrease with increased trimming, there is a concomitant decrease in the number of matching paired reads mapped, as well as the number of ORFs that can be identified [7]. Second, the number of distinct transcripts detected through *de novo* assembly decreases with more stringent trimming

[8]. Finally, trimming can increase the number of false positive variant calls in genome sequencing studies [9]. All of these findings are consistent with increasing difficulty in unambiguously aligning shorter reads to a reference genome and/or reconstructing less total sequence into longer contiguous sequences.

The above studies have all investigated the influence of trimming on the immediately downstream steps of read alignment and transcriptome reconstruction [6–9], but it remains to be determined how trimming impacts further downstream analyses – for example, expression estimation and statistical identification of differentially expressed genes. One might expect that the specificity of read alignments could impact gene expression estimates and have vital effects on differential expression predictions. Consistent with this possibility, removing the first ten bases from all reads, irrespective of quality scores, led to an approximately two percent decrease in the number of differentially expressed genes detected in the *D. melanogaster* larval central nervous system following neuronal knockdown of a factor involved in spliceosome assembly [10]. More generally, one might expect that aggressive quality-based trimming would decrease the likelihood of detecting false positives that arise from erroneous mapping due to sequencing errors, while simultaneously reducing the sensitivity of detecting differentially expressed genes, since expression estimates would have reduced precision as a consequence of less sequencing information contributing to their measurement.

Here, we set out to explore the effects of quality-based trimming on gene expression analysis and report that multiple forms of bias in gene and isoform expression levels are apparent when comparing an untrimmed RNA-Seq data set to the same data set with trimming applied. Most of this bias can be removed by imposing a minimum read length requirement following trimming, suggesting that the gains in base calling accuracy that result from aggressive trimming are offset by the detrimental effects of estimating gene expression from short reads. However, despite the ability to correct much of the short read-associated bias by imposing a minimum length filter, a subset of biased genes remains resistant to correction. Thus, we caution that aggressive trimming of RNA-Seq data can introduce bias and unpredictability into RNA-Seq gene expression estimates, which can subsequently impact downstream differential expression analysis.

2.3 Methods

Fly stocks

The following lines were used in this study: *Gal4*$^{21\text{-}7}$ [32], *UAS-RedStinger* [33], *UAS-mCD8GFP* [34].

Flow cytometry

Third instar larvae were filleted by microdissection in PBS. Internal organs and thoracic segments were removed, and the remaining body walls were digested in 500 µl 0.9 mg/ml (200 U/ml) collagenase in PBS for 18 min at 37°C with mechanical agitation (1000 rpm on a 3 mm orbit diameter shaker, with trituration every 6 minutes). Debris was removed by filtering cell suspensions through a 70 µm nylon filter, and cells were isolated to high purity using two successive rounds of sorting on a FACSAria II (BD Biosciences, San Jose, CA). Four samples of 100 cells each were captured into 2 µl of SMARTer lysis mix (described below) and were immediately processed for RNA-Seq.

RNA-Seq

Total RNA from lysed cells was converted to pre-amplified cDNA libraries using template-switching reverse transcription [35, 36] as implemented in the SMARTer Ultra-low input kit (Clontech, Mountain View, CA), but with modified procedures for low cell number analysis (Fluidigm, South San Francisco, CA). Pre-amplified cDNA libraries were diluted to 0.25 ng/ul. Fragmentation was performed enzymatically using a Nextera XT DNA kit (Illumina, San Diego, CA), and barcoded samples were multiplexed, pooled, purified using Agencourt AMPure XP beads (Beckman Coulter Genomics, Danvers, MA), and quality controlled on a Bioanalyzer 2100 using a high sensitivity dsDNA assay (Agilent Technologies, Santa Clara, CA). Quality-controlled libraries were sequenced as 51 base single end reads on a HiSeq 2500 running in high-output mode at the UCSF Center for Advanced Technology (San Francisco, CA). Reads were demultiplexed with CASAVA (Illumina), and read quality was assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). One library was sequenced twice in order to increase sequencing depth. In total, the four replicate samples were comprised of 7, 13, 14, and 21 million reads passing sequencing filters.

Trimming with SolexaQA

Trimming was performed with SolexaQA version 3.1.2 [14], which scans for the longest contiguous run in the sequence with quality scores at or above the user-provided value. To perform filtering on read lengths, the lengthsort command was run following the initial trimming command. Example commands for these and all other tools can be found in Additional File 7.

Trimming with Trimmommatic

Trimming was performed with Trimmomatic version 0.33 [15]. We used the quality filtering functionality of this tool with a sliding window, which scans through reads from the 5' end, and removes following bases from the 3' end once the average quality score within the window drops below a user-specified value.

Trimming with ConDeTri

Trimming was performed with ConDeTri version 2.2 [16]. For each instance, both a high quality and a low quality score were provided as parameters; the low quality scores were held either five or ten below the high quality scores for all combinations tested. Briefly, ConDeTri removes bases from the 3' end of reads that are below the high quality score. Once a base is encountered that surpasses the high quality score, bases are retained so long as the bases between the low quality score and high quality score, as a fraction of total bases, does not rise above a default threshold of 0.2. All bases distal to a base below the low quality threshold are discarded. Aside from the quality scores, the only other parameter that was altered from the defaults was the minimum length, which was removed rather than using the default value of 50 to accommodate the 51base sequencing reads used in this study.

Alignment to the transcriptome

After trimming, reads were aligned to the *D. melanogaster* genome, FlyBase genome release 6.04, to the *Rattus norvegicus* genome, Ensembl release 5.0, or to the *Saccharomyces cerevisiae* genome, Ensembl release R64-1-1. TopHat2 version 2.0.14 [17] and Bowtie2 version 2.2.3 [17, 37] were used for alignment using two threads, but otherwise with all default parameters. The aligned reads, alignment summary, and junction alignment files were used in further analysis. In addition to the above, several other

alignment/expression estimation approaches were employed. In one case, gene-level counts from the TopHat2 output were determined using HTSeq version 0.6.0 [38]. All standard parameters were used in the gene counts mode for the aligner STAR version 2.4.2a [24]. RSEM version 1.2.22 [25] was used in combination with STAR version 2.4.2a [24].

Gene expression analysis

Differential gene expression analysis was performed using Cuffdiff2 version 2.2.1 [20]. In each case, the three (yeast) or four (fly, rat) trimmed samples were compared to the three or four samples without any trimming. A reference transcriptome was provided, and as such any novel junctions detected by TopHat2 were not modeled. All other parameters were their default. The gene_exp.diff and isoform_exp.diff output files were used to determine the significantly differentially expressed genes and isoforms as well as expression values in both trimmed and untrimmed samples. For diverse pipeline analysis, differential gene expression analysis on counts data was performed using the R package DESeq2 version 1.10.0 [26] or the R package EdgeR version 3.13.4 [27].

Gene and isoform parameter analysis

Gene and isoform parameters were generated from the Cuffdiff2 output (gene expression) and the FlyBase release 6.04 transcriptome (isoform length, number of exons per isoform). Significance in comparisons of these parameters was assessed using a Mann-Whitney U test. The number of genes to which multi-hit reads mapped was determined by identifying multi-hits using the TopHat2 output, followed by using these reads as input to Cufflinks version 2.2.1 [19]. All genes which showed non-zero expression from any of the four multi-hit samples were considered to be a target of multi-hit reads. Significance was assessed using a Poisson test. GC content and Markov entropy scores were calculated as previously described [22, 39] using a publicly available Perl package (https://github.com/caballero/SeqComplex.git). Significance was assessed using a two-tailed Student *t* test assuming unequal variances. An adjusted p-value of 0.05 after Benjamini-Hochberg correction was deemed significant.

Correlations with microarray expression data

Microarray intensity values were retrieved from the NCBI Gene Expression Omnibus (GEO) with the R package GEOquery version 2.37 (https://github.com/seandavi/GEOquery). Probes were mapped to the same genome to which RNA-seq reads were aligned, and any probes mapping to more than one gene were discarded. The normalized intensity values were averaged across all samples and all probes mapping to each gene to calculate gene-level intensity values. Pearson's correlations were used to measure the correlation between the average gene expression based on microarray intensity data and the estimated gene expression based on RNA-Seq data, after imposing a lower expression cutoff of 1 FPKM.

2.4 Results and Discussion

2.4a *Quality-based trimming of ultralow-input RNA-Seq data increases mappability*

Previous work has shown that quality-based trimming of RNA-Seq data can lead to greatly increased mappability of reads (*i.e.*, percentage of input reads that can be successfully aligned to a genome) [6]. However, this increased mappability of reads remaining after trimming comes at the expense of a dramatic reduction in the absolute number of aligned reads, as a consequence of some reads failing to pass minimum quality criteria during trimming. We predicted that this loss of information would impact analyses downstream of alignment; in particular, gene expression estimation. To assess this, we first generated RNA-Seq data from multi-dendritic (md) sensory neurons from *D. melanogaster* larvae, which had not yet been transcriptionally characterized by RNA-Seq despite their frequent use as a model system for neuronal development [11]. This approach was selected over those based on cells grown in culture to maximize physiological relevance. In this regard, the influence of trimming on expression measurement is particularly relevant to approaches using RNA-Seq for systematic identification of cell type in the nervous system [12, 13]. Neurons were sorted to high purity using two consecutive rounds of flow cytometry (Figure 2.1a, b) and four samples comprised of 100 cells each were processed by SMART-Seq and sequenced on a HiSeq 2500. Each sample comprised at least seven million unpaired 51 base reads and was of high overall quality (Figure 2.1c).

To assess whether trimming improved mappability of our samples, as has been reported elsewhere [6], we trimmed our sensory neuron data sets with three different trimming algorithms and determined mappability. First, we evaluated SolexaQA, a sliding window trimmer that offers a balanced tradeoff between mappability and the number of mapped reads [6, 14]. We also evaluated Trimmomatic, which was shown to achieve high mappability with less aggressive trimming [6, 15], and ConDeTri, which demonstrated high mappability when used aggressively [6, 16]. We varied the quality score threshold from 10, corresponding to a 1 in 10 chance of an incorrect base, up to 40, corresponding to a 1 in 10,000 chance of an incorrect base – the highest confidence score assigned in Illumina sequencing data. After trimming, data were aligned to the annotated *D. melanogaster* transcriptome using TopHat2 [17]. As previously shown with another high quality RNA-Seq data set [6], mappability increased with increasing quality requirements, but the absolute number of aligned reads decreased (Figure 2.1, SolexaQA; Additional File 1, Trimmomatic and ConDeTri). Thus, the impact of trimming on the mappability of the high quality reads generated from the small cell numbers employed in our study was similar to that observed from samples generated from abundant input RNA [6].

2.4b Junction spanning reads decrease disproportionately following trimming

Although trimming increases overall mappability, it can also substantially shorten many reads, depending on the aggressiveness of the trimming parameters. We reasoned that this reduction in information content might introduce one or more forms of bias during read alignment. In particular, we predicted that there would be a disproportionate bias against reads aligning to exon-exon junctions, since alignment to such sites requires sequences long enough to span both the splice donor and acceptor sides of the junction. TopHat2 requires that reads, either singly or in combination with other reads, align for at least eight contiguous bases with no mismatches on both sides of a junction for initial junction detection, though subsequent reads may span a shorter distance and will still map to the junction [17, 18]. This is in contrast to aligning to non-junction locations, which minimally requires twelve contiguous bases with no more than one mismatch. As predicted, we observe that trimming disproportionately decreases the proportion of reads mapped to exon-exon junctions. The frequency of reads aligned to junctions, as a function of the total number of reads aligned, decreases as trimming quality score threshold increases,

from 8.5% (4.27 million reads aligned to junctions per 50.34 million total reads aligned in all samples combined) without trimming to 3.0% (0.14 million reads per 4.54 million total reads) at Q40 (Figure 2.3a, b). Interestingly, this is not the case with the frequency at which junctions are *detected*, as the number of junctions detected per reads aligned increases with increasing quality score stringency, from 1.5 junctions detected per thousand reads mapped without trimming (74 thousand junctions detected) to 4.3 junctions detected per thousand reads mapped at Q40 (20 thousand junctions detected) (Figure 2.3c, d). Although the reason for this is unclear, we speculate that at the read coverage depth in our data, our ability to detect junctions is not constrained by coverage even after trimming, resulting in the increased frequency of junction detection largely being driven by the decrease in the total number of aligned reads.

2.4c Bias in expression levels estimated from untrimmed and trimmed reads

We predicted that the decreased frequency of reads aligning to junctions would change estimates of isoform expression levels, since accurate alignment of reads to junctions contributes to the assignment of reads to specific isoforms [19]. Such bias would be expected to manifest as significantly different expression between trimmed and untrimmed samples, which we tested using Cuffdiff2 [20]. We note that throughout this work we refer to bias in the sense that gene expression is different between the groups, but with limited *a priori* knowledge of whether the gene expression estimates based on untrimmed or trimmed reads are more accurately reflective of the true expression levels (discussed in more detail below).

As predicted, the expression of many isoforms was significantly altered by quality score trimming, with hundreds of differentially expressed isoforms detected with aggressive trimming (Figure 2.4a, b). This finding holds even if novel junction discovery, the default behavior of TopHat2 mapping, is disabled (Additional File 2), since only the annotated transcriptome and junctions were used for modeling by Cuffdiff2. Because Cuffdiff2 estimates gene-level expression as the sum of the expression of all individual isoforms [19], we further predicted that in addition to isoforms, genes would exhibit expression bias following trimming. As expected, we observed a progressively increasing number of significant differentially expressed genes between our untrimmed data set and trimmed data sets with increasingly aggressive quality filtering (Figure 2.4a, c). At the most stringent quality score, Q40, Cuffdiff2 identified

1829 genes, representing 10.5% of all annotated genes, biased towards higher expression in either the untrimmed or trimmed data set, suggesting that trimming can have a substantial effect on the apparent composition of a sample.

Although the junction-alignment bias described above might play a role in these differential expression estimates, other factors must contribute as well since junction bias alone was insufficient to explain all of the observed bias. For example, we found that loss of junction reads did not uni-directionally decrease expression estimates. Instead, bias toward higher expression in untrimmed data was detected for some isoforms, but toward higher expression in trimmed data for others, including comparisons in which the number of junctions was held constant. Low expression level was also not a primary factor driving significance—no significant genes or isoforms exhibited expression values, measured as fragments per kilobase of transcript per million mapped reads (FPKM), of less than one in both the untrimmed and trimmed data sets (Figure 2.4). Thus, it is likely that trimming introduces or corrects multiple sources of bias in gene expression estimation, relative to untrimmed reads, and that filtering based on expression level would not provide a means by which to eliminate this bias.

2.4d Short trimmed reads are the predominant source of bias

Since bias resulting from differential alignment of junction-spanning reads could not fully account for the observed differences in expression estimated from untrimmed and trimmed reads, we next hypothesized that read length might contribute to the observed bias through other mechanisms. In addition to removing reads of very low quality in their entirety, trimming also shortens reads of mixed quality to preserve only high quality bases. Thus, the trimmed data sets have a distribution of read lengths as compared to the uniform read length in the untrimmed data set (Figure 2.5a, Additional File 3). We predicted that shorter reads would align to more locations than longer reads, and that this promiscuity in mapping would drive some of the observed differential expression estimates. To evaluate this, we removed all reads below a fixed length in the most heavily trimmed SolexaQA data set, Q40-trimmed, and compared gene expression between these data and untrimmed reads. Minimum length requirements below 12 bases had no effect on the number of differentially expressed genes or isoforms identified by Cuffdiff2, which was expected since such reads fall below the default threshold for reads that TopHat2

attempts to align. However, following length filtering using longer thresholds, much of the bias both in isoform and gene expression between untrimmed and trimmed samples was eliminated (Figure 2.5b-d). At the highest quality score, Q40, the number of significantly biased genes was reduced from 1829 to 150 and the number of significantly biased isoforms was reduced from 1269 to 41 when the minimum read length was increased from 1 to 36. Increasing stringency beyond a minimum of length of 36 was not attempted because few Q40-trimmed reads exceeded this length.

The impact of short reads on trimming-induced bias was corroborated by results from trimming with Trimmomatic and ConDeTri. Rather than searching sequencing reads for the longest run of bases over a given quality, both of these trimmers search from the end of reads, such that if a stretch of high quality is encountered near one of the ends, only the bases outside of that run will be truncated. One consequence of this approach to trimming is that the retained reads are considerably longer, with very few short reads retained as compared with SolexaQA (see Additional File 3). Consistent with the hypothesis that read length drives bias, even fairly aggressive application of these trimmers results in considerably less bias than trimming with SolexaQA, with a maximum of 9 biased genes with Trimmomatic (q = 30) and 28 biased genes with ConDeTri (hq = 39, lq = 34). Thus, short reads generated upon trimming are an important driver of bias in gene expression estimates, but this can be partially offset by imposing stringent minimum length filters.

Finally, we note that the long reads that remain after both stringent quality-based trimming and length filtering can be mapped with high accuracy; over 97% of 36-mers present in the *D. melanogaster* genome are unique. Given that bias is minimized between the full, untrimmed data set and this aggressively trimmed and length filtered high confidence data set, this suggests that the full, untrimmed data set generates a more faithful representation of true gene expression estimates than those derived from aggressively trimmed data containing short reads.

2.4e Additional factors contribute to gene expression bias

Although imposing read length requirements counteracted bias introduced by trimming, notable differences remained between the untrimmed and the processed data, and we next sought to identify additional drivers that could account for the residual bias. We divided the genes and isoforms differentially

expressed at Q40 without length filtering into two groups—correctable and resistant—according to whether or not expression bias could be corrected by length filtering (minimum length = 36), as assessed using Cuffdiff2.

We assessed five parameters related to read alignment and transcript structure of the biased genes and isoforms. We hypothesized that poorly expressed genes would be more strongly impacted by promiscuous alignment of short reads than highly expressed genes, due to the proportion of inappropriately aligning reads being higher for poorly expressed genes. Consistent with this prediction, the expression levels of resistant genes and correctable genes differed prior to length filtering, with the resistant genes exhibiting a median expression of 56 FPKM, as compared with a median expression of 28 FPKM among the biased genes corrected by length filtering ($p<0.05$, Mann-Whitney test) (Figure 2.6a).

Because short reads are more likely to map to multiple locations in the genome (referred to as "multi-hits" for consistency with TopHat2 nomenclature), we next investigated how this property is associated with the observed biases. Before length filtering, multi-hit reads mapped to over 99% of detected genes, indicating that expression estimates were broadly influenced by short reads aligning to multiple locations. However, this was not the case after imposing a minimum read length requirement of 36 bases: after filtering, 10% of genes resistant to bias-correction, but only 1.8% of correctable genes, contained any multi-hit reads ($p<0.05$, Poisson test). Thus, mapping of non-unique short reads is rampant in aggressively trimmed data, and may continue to contribute a small portion of the residual bias even after length filtering. To more directly assess the role of multi-hits in differential expression following trimming, we repeated differential expression analysis using only uniquely mapping reads. Eliminating multi-hit reads greatly reduced the number of differentially expressed genes and isoforms after trimming at Q40 to 75 and 61, respectively (Additional File 4). However, as would be predicted based on the low percentage of non-unique reads present after length filtering, the effect on differential expression following length filtering was minimal (see Additional File 4), suggesting that multi-hits are not the primary driver of the residual bias after length filtering, and that additional factors may play a role. Although these data indicate that gene expression estimation from trimmed reads is stabilized by excluding multi-hits, others have found that allowing multi-hits increases the accuracy of expression estimates from 36-base RNA-Seq reads [21]. Thus, exclusion of all multi-hits could introduce bias as well; whether this bias or

that associated with promiscuous alignment of short reads is more tolerable will need to be evaluated on a case-by-case basis.

The ability of short reads to align to multiple locations might be influenced by the intrinsic sequence content of a given gene or isoform. Specifically, we predicted that bias-correctable genes might exhibit lower sequence complexity, which would result in higher rates of multi-hit mapping, but that could be corrected by length filtering. To examine sequence complexity, we assessed entropy of isoform sequences in the two groups using Markov models for 1 to 6 base pair oligonucleotides [22]. Two of the six measures of complexity were significantly different between the correctable and resistant groups, with the correctable group exhibiting lower complexity in both cases as predicted (Additional File 5). However, we also noted that length filtering-resistant isoforms exhibited significantly higher GC content (Figure 2.6b), and that both of the significant complexity measures were also significantly correlated with GC content. This observation suggested that GC content, rather than complexity per se, might be the primary underlying factor driving resistance to correction by length filtering. Notably, genes with high GC content exhibit disproportionately high expression values in RNA-Seq studies [23], which is also consistent with our observation that FPKM is associated with resistance to bias-correction (Figure 6a). In anticipation of this potential bias, Cuffdiff2 was run with the optional fragment bias correction protocol [19] enabled; however, as evidenced by the above findings, some GC content bias remained.

We next evaluated structural properties of transcript isoforms—specifically, isoform length and number of exons—as a source of resistance to bias-correction through length filtering. The distributions of transcript lengths were not different between the two groups ($p>0.05$, Mann-Whitney test) (Figure 2.6c). In contrast, the number of exons, and therefore also the number of junctions, was higher in the correctable group (4.7 exons per isoform) as compared with the resistant group (3.2 exons per isoform) (Figure 2.6d) ($p<0.05$, Mann-Whitney test). In addition, both the frequency of junction detection and frequency of reads mapped to junctions increased with increasingly stringent length filtering (Additional File 6). Together, these data suggest that length filtering of quality-filtered data improves detection of exon-exon junctions in addition to reducing spurious multi-hit alignments.

2.4f Trimming-induced differential expression is manifest in diverse analysis pipelines

Although TopHat2 and Cufflinks2 are widely used for analyzing RNA-Seq data, alternative tools have been gaining broad acceptance. Different tools vary in their underlying assumptions about read distribution and in their approach to handling non-uniquely mapping reads; therefore, we next examined whether the trimming-induced biases we identified are generalizable to other pipelines. Most tools assess differential expression based on gene-level counts, without discrimination of isoforms; thus, we focused our analysis on differential gene expression. We implemented four additional pipelines using the read aligners STAR [24] and RSEM [25] in combination with the differential analysis tools DESeq2 [26] and EdgeR [27]. Consistent with our TopHat2 / Cufflinks2 results, significantly differentially expressed genes were detected with each additional pipeline following trimming with stringent quality parameters (SolexaQA with Q = 40), albeit fewer than our original analysis identified, and these largely disappeared when a length filter was imposed (Table 2.1). The differences in the number of differentially expressed genes between analysis tools may be due to inherent differences in how liberal or conservative the programs are in calling significant differences, as previously reported [28, 29]. Despite differences in the scale of the effect, all of these tools indicated that trimming affects gene expression estimates in this *D. melanogaster* RNA-Seq data set.

2.4g Trimming-induced differential expression is manifest in diverse RNA-Seq data sets

We next assessed whether the effects of trimming found in the *D. melanogaster* RNA-Seq data set were observed with other independently generated RNA-Seq data. For these analyses, we chose data sets derived from different organisms (rat livers [30], yeast cultures [31]) and generated in different labs using different library preparation and sequencing protocols. These additional data sets were comprised of samples with paired 101 base reads; thus, we anticipated that the negative effects of trimming would be less severe since longer reads are less likely to map to multiple locations, and paired reads must map concordantly. Instead, we found that trimming had a more pronounced effect on these data than on our original data (Figure 2.7). Using the SolexaQA / TopHat2 / Cuffdiff2 pipeline, we found that 54% of genes (14,470 of 26,689 total) in the rat sample and 78% of genes (5552 of 7126 total) in the yeast sample were significantly altered in their expression when the most aggressive trimming, Q40, was

applied. As in the *D. melanogaster* data set, imposing a minimum length filter of 36 bases substantially reduced the number of differentially expressed genes, down to 2% (rat) and 10% (yeast) of all genes. We note that smaller fold changes between the trimmed and untrimmed samples were called as significantly different (visualized as points close to the identity line in Figure 2.7a, b) in these two data sets than in the original data set, which might be due to lower variance between replicate samples and/or increased accuracy in alignments due to the use of paired reads. Thus, we expect that quality-based trimming will alter gene and isoform level expression estimates across RNA-Seq data sets, though the extent to which estimates change will depend on characteristics specific to each data set.

2.4h Aggressive trimming decreases concordance with microarray expression estimates

Given that trimming causes substantial changes in gene expression estimates across multiple RNA-Seq data sets, we next investigated whether trimming improved or reduced the accuracy of expression estimates. As a biological standard for gene expression, we used the rat and yeast data sets described above, in which the same RNA libraries were subjected to genome-wide gene expression analysis both by RNA-Seq and by hybridization to microarrays [30, 31]. Specifically, we reasoned that if trimming reduced the accuracy of RNA-Seq based expression estimates, we should observe decreased concordance between the RNA-Seq and microarray expression values in trimmed RNA-Seq data sets. This is precisely what we observed (Table 2.2). In the two independent RNA-Seq/microarray data sets, expression estimates from untrimmed RNA-Seq data were most highly correlated with microarray expression estimates, though even moderately aggressive trimming, up to Q30, minimally reduced these correlations. By contrast, aggressive trimming led to substantially reduced correlations with microarray data. Length filtering slightly improved the correlations with microarray estimates for the heavily trimmed rat data; however, length filtering of the yeast data further decreased correlations, suggesting that this additional filtering may not universally counteract trimming-induced bias. Thus, by validation with an independent technique, we conclude that no or low trimming thresholds are most likely to result in the highest accuracy for RNA-Seq based expression estimates.

2.5 Conclusions

The data we present here provide evidence that aggressive quality-based trimming can strongly influence estimation of gene and isoform expression levels, which subsequently impacts identification of differentially expressed genes. A considerable source of the observed differences can be attributed to the alignment of shorter reads that result from trimming. Imposing minimum read length requirements reverts gene expression estimates to values closer to estimates produced from untrimmed reads, suggesting that untrimmed or trimmed, length-filtered reads—the latter of which likely represent the highest quality reads within a data set—may most accurately reflect the actual library composition.

Because different experiments have different goals, individual researchers must determine whether or not trimming will be beneficial for their particular application. For example, in genome sequencing or for RNA-Seq experiments where extremely large numbers of reads are available, modest trimming may provide benefits. Further, in data sets with low average base calling quality, or in library preparation protocols that are susceptible to adapter contamination, trimming may allow the recovery of reads which would otherwise be detrimental to expression estimation. Both of these attributes were more common in early RNA-Seq studies, so trimming may be particularly useful when re-analyzing such data. One potential improvement may be to use longer sequencing reads, such as 100 or 150 bases, so that longer reads remain after trimming low quality bases from either end, though our results demonstrate that this alone will not prevent the introduction of trimming-induced expression changes. However, we re-iterate previously voiced concerns [7, 8] that mappability should not be used as the sole criterion for performance. Furthermore, our results suggest that aggressive trimming adversely affects the accuracy of expression estimates. Therefore, if trimming is applied, extreme care should be used, and other measures such as length filtering should be considered in the pre-processing pipeline to minimize the introduction of unwanted bias.

2.6 Declarations

Availability of Supporting Data

The fly data set generated in this article is available in the NCBI Sequence Read Archive (SRA) and in the Gene Expression Omnibus (GEO) under accession number GSE72884. The rat RNA-Seq data sets

used were obtained from the SRA under accession numbers SRR1178065, SRR1178067, SRR1178068, and SRR1178069 and the corresponding microarray data sets were obtained from GEO under accession numbers GSM116428, GSM1161435, GSM1161439, and GSM1161443. The yeast RNA-Seq data sets used were obtained from the SRA under accession numbers SRR453569, SRR453570, and SRR453571, and the corresponding microarray data sets were obtained from GEO under accession numbers GSM923093, GSM923094, and GSM923095.

List of Abbreviations

FPKM: Fragments per kilobase of transcript per million reads mapped

Q: Phred quality score

RNA-Seq: RNA sequencing

Competing Interests

The authors declare no competing interests.

Authors' Contributions

CRW, JZP, and CCK conceived and designed the study. CRW, JZP, and CCK prepared RNA-Seq libraries. CRW and CCK conducted trimming analysis. CRW and AB conducted the length filtering analysis. CRW conducted the paired RNA-seq/microarray expression correlation analysis. CRW, JZP, and CCK wrote the manuscript.

Acknowledgements

We thank the Bloomington *Drosophila* Stock Center for fly stocks. This work was supported by NIAID P30 AI027763 and NIDDK P30 DK063720 to CCK; NINDS R01-NS076614, a March of Dimes Basil O'Connor Starter Scholar Award, a Klingenstein Fellowship in Neuroscience, a UW Royalty Research Award and a UW Research Innovation award to JZP; an ACCMA Community Health Foundation Summer Stipend and a Schoeneman Scholarship to AB; and an NSF Graduate Research Fellowship (DGE-1256032) to CRW.

2.7 Supplemental Information

Additional files referenced in the text of this chapter are available with the published version of this manuscript, found at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0956-2. The content of these files is as follows:

Additional File 1. Influence of trimming with Trimmomatic and ConDeTri on mappability. (a) The total number of input reads (light bars) and reads aligned to the transcriptome (dark bars) from four RNA-Seq data sets trimmed at a range of quality scores with Trimmomatic. Q scores are 5 apart from 5 to 25, and every Q score from 25 to 34 is shown. No reads survived at or above a Q score of 35. (b) The mappability, or number of aligned reads per total input reads, per sample trimmed with Trimmomatic. (c) The total number of input reads (light bars) and reads aligned to the transcriptome (dark bars) from four RNA-Seq data sets trimmed with ConDeTri. The high quality (HQ) score for trimming is indicated under the first of each pair of bars, and a low quality (LQ) score five (red bars) or ten (orange bars) below was used. (d) The mappability per sample trimmed with ConDeTri. Input reads shorter than 12 bases were not included in the mappability calculations, as these are discarded by TopHat prior to alignment. Error bars represent standard deviations.

Additional File 2. Influence of novel junction discovery on isoform and gene expression levels. Comparison of the expression estimates of isoforms and genes between the SolexaQA Q40-trimmed and the untrimmed data set, after aligning reads to the transcriptome using TopHat2 with novel junction discovery disabled. Red dots represent statistically significant differential expression between data sets.

Additional File 3. Distribution of read lengths after trimming. Density plots show the distributions of read lengths at multiple Q scores following trimming with SolexaQA (a), Trimmomatic (b), and ConDeTri (c).

Additional File 4. Influence of multi-hits on isoform and gene expression levels. Comparison of the expression estimates of isoforms and genes between the SolexaQA Q40-trimmed without (a) or with (b) a

minimum length requirement and the untrimmed data set, after aligning reads to the transcriptome using TopHat2 with multi-hits excluded. Red dots represent statistically significant differential expression between data sets.

Additional File 5. Relationship between length-filtering resistant bias and sequence complexity measures. The distribution of complexity scores for length filtering-correctable and -resistant isoforms, assessed with a Markov model for entropy of oligonucleotides of length one (a), two (b), three (c), four (d), five (e), or six (f). *, $p<0.05$ following Benjamini-Hochberg adjustment. Bars represent the mean. For clarity, not all data points are depicted. Cor, correctable. Res, resistant.

Additional File 6. Influence of minimum length requirements on junction alignment and detection. (a) The average number of reads aligned to junctions per sample with increasing minimum read length requirements after trimming with SolexaQA, Q=40. (b) The average frequency of reads aligned to junctions (number of reads aligned to junctions per total reads aligned). (c) The average number of junctions detected per sample. (d) The average frequency of junction detection (number of junctions detected per total reads mapped). For all panels, data were normalized to the Q40 value with no minimum length filter, on a per sample basis. Error bars represent standard deviations.

Additional File 7. Example commands used for analysis tools. Table showing example commands for all analysis tools used.

2.8 References

1. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV: **The challenges of sequencing by synthesis**. *Nat Biotechnol* 2009, **27**:1013–1023.

2. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming**. *Nucleic Acids Res* 2010, **38**:e131.

3. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, **36**:e105.

4. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**:186–194.

5. Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, Sivils KL, Olson LE, Iwayama T, Freeman WM, Lessard CJ, Wren JD: **Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data**. *BMC Bioinformatics* 2015, **16 Suppl 13**:S10.

6. Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM: **An extensive evaluation of read trimming effects on Illumina NGS data analysis**. *PloS One* 2013, **8**:e85024.

7. MacManes MD: **On the optimal trimming of high-throughput mRNA sequence data**. *Front Genet* 2014, **5**:13.

8. Mbandi SK, Hesse U, Rees DJG, Christoffels A: **A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads**. *Front Genet* 2014, **5**:17.

9. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y: **Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data**. *BMC Genomics* 2012, **13 Suppl 8**:S8.

10. Amaral AJ, Brito FF, Chobanyan T, Yoshikawa S, Yokokura T, Van Vactor D, Gama-Carvalho M: **Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models**. *Front Genet* 2014, **5**:43.

11. Singhania A, Grueber WB: **Development of the embryonic and larval peripheral nervous system of Drosophila**. *Wiley Interdiscip Rev Dev Biol* 2014, **3**:193–210.

12. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S, Phatnani HP, Guarnieri P, Caneda C, Ruderisch N, Deng S, Liddelow SA, Zhang C, Daneman R, Maniatis T, Barres BA, Wu JQ: **An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex**. *J Neurosci Off J Soc Neurosci* 2014, **34**:11929–11947.

13. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, Hjerling-Leffler J, Haeggström J, Kharchenko O, Kharchenko PV, Linnarsson S, Ernfors P: **Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing**. *Nat Neurosci* 2015, **18**:145–153.

14. Cox MP, Peterson DA, Biggs PJ: **SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data**. *BMC Bioinformatics* 2010, **11**:485.

15. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinforma Oxf Engl* 2014, **30**:2114–2120.

16. Smeds L, Künstner A: **ConDeTri--a content dependent read trimmer for Illumina data**. *PloS One* 2011, **6**:e26314.

17. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**:R36.

18. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinforma Oxf Engl* 2009, **25**:1105–1111.

19. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**:562–578.

20. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with RNA-seq**. *Nat Biotechnol* 2013, **31**:46–53.

21. Odawara J, Harada A, Yoshimi T, Maehara K, Tachibana T, Okada S, Akashi K, Ohkawa Y: **The classification of mRNA expression levels by the phosphorylation state of RNAPII CTD based on a combined genome-wide approach**. *BMC Genomics* 2011, **12**:516.

22. Orlov YL, Potapov VN: **Complexity: an internet resource for analysis of DNA sequence complexity**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W628–633.

23. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-Seq data**. *BMC Bioinformatics* 2011, **12**:480.

24. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinforma Oxf Engl* 2013, **29**:15–21.

25. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.

26. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**:550.

27. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinforma Oxf Engl* 2010, **26**:139–140.

28. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, Zhao Q-Y: **A comparative study of techniques for differential expression analysis on RNA-Seq data**. *PloS One* 2014, **9**:e103207.

29. Seyednasrollah F, Laiho A, Elo LL: **Comparison of software packages for detecting differential expression in RNA-seq studies**. *Brief Bioinform* 2015, **16**:59–70.

30. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Łabaj PP, Kreil DP, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian H-R, Lancashire LJ, Bessarabova M, Nikolsky Y, Furlanello C, et al.: **The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance**. *Nat Biotechnol* 2014, **32**:926–932.

31. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J: **A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae**. *Nucleic Acids Res* 2012, **40**:10084–10097.

32. Song W, Onishi M, Jan LY, Jan YN: **Peripheral multidendritic sensory neurons are necessary for rhythmic locomotion behavior in Drosophila larvae**. *Proc Natl Acad Sci U S A* 2007, **104**:5199–5204.

33. Barolo S, Castro B, Posakony JW: **New Drosophila transgenic reporters: insulated P-element vectors expressing fast-maturing RFP**. *BioTechniques* 2004, **36**:436–440, 442.

34. Lee T, Luo L: **Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis**. *Neuron* 1999, **22**:451–461.

35. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, Chenchik A: **Amplification of cDNA ends based on template-switching effect and step-out PCR**. *Nucleic Acids Res* 1999, **27**:1558–1560.

36. Petalidis L, Bhattacharyya S, Morris GA, Collins VP, Freeman TC, Lyons PA: **Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis**. *Nucleic Acids Res* 2003, **31**:e142.

37. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**:357–359.

38. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data**. *Bioinforma Oxf Engl* 2015, **31**:166–169.

39. Caballero J, Smit AFA, Hood L, Glusman G: **Realistic artificial DNA sequences as negative controls for computational genomics**. *Nucleic Acids Res* 2014, **42**:e99.

2.9 Chapter Two Figures

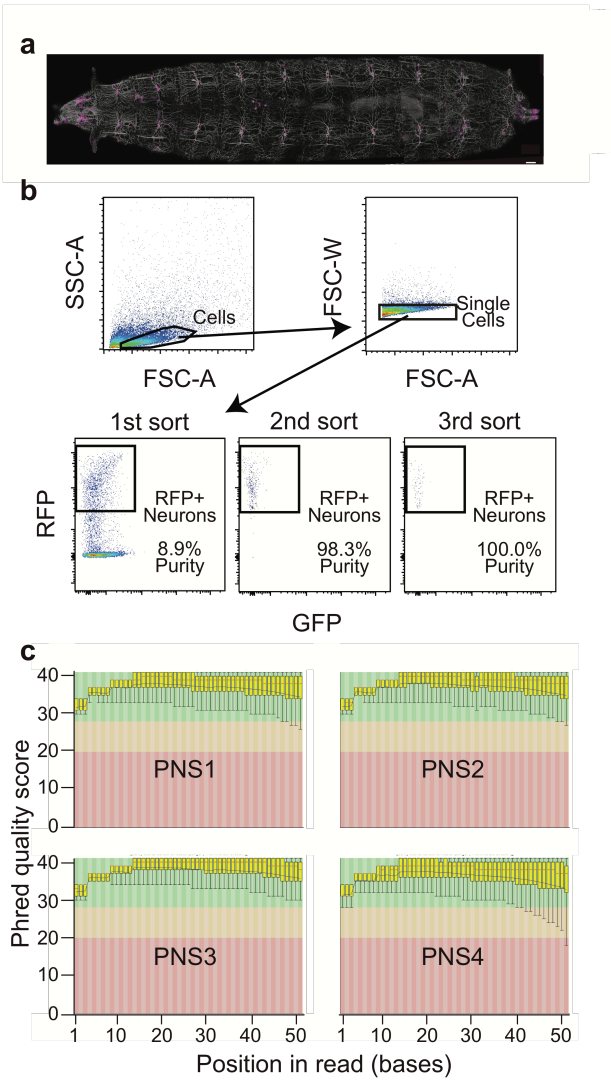

Figure 2.1 High quality RNA-Seq data generated from D. melanogaster sensory neurons

(a) Confocal image of *Drosophila* larval sensory neurons expressing a nuclear-targeted version of mRFP (magenta) and a membrane-targeted version of GFP (white). Genotype: w^{118}; $Gal4^{21\text{-}7}$, *UAS-mCD8-GFP/UAS-Red-Stinger*. Scale bar is 100μm. (b) Representative flow plots of *D. melanogaster* neurons. Plots show three progressive gates to identify RFP^+ neurons, followed by two additional re-sorts with the same gates to assess purity. Compensated fluorescent values are shown. (c) Box plots generated in FastQC show average (blue lines) and median (red lines) quality scores across all read positions for each of the four independent replicate samples.

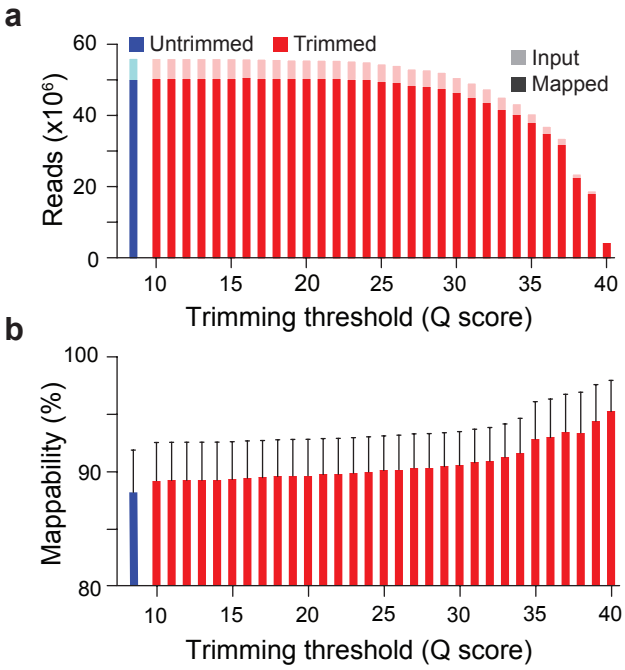

Figure 2.2 Influence of quality-based trimming on mappability

(a) The total number of input reads (light bars) and reads aligned to the transcriptome (dark bars) from four RNA-Seq data sets trimmed at a range of quality scores with SolexaQA. (b) The mappability, or number of aligned reads per total input reads per sample. Input reads shorter than 12 bases were not included in the calculation, as these are discarded by TopHat2 prior to alignment. Error bars represent standard deviations.

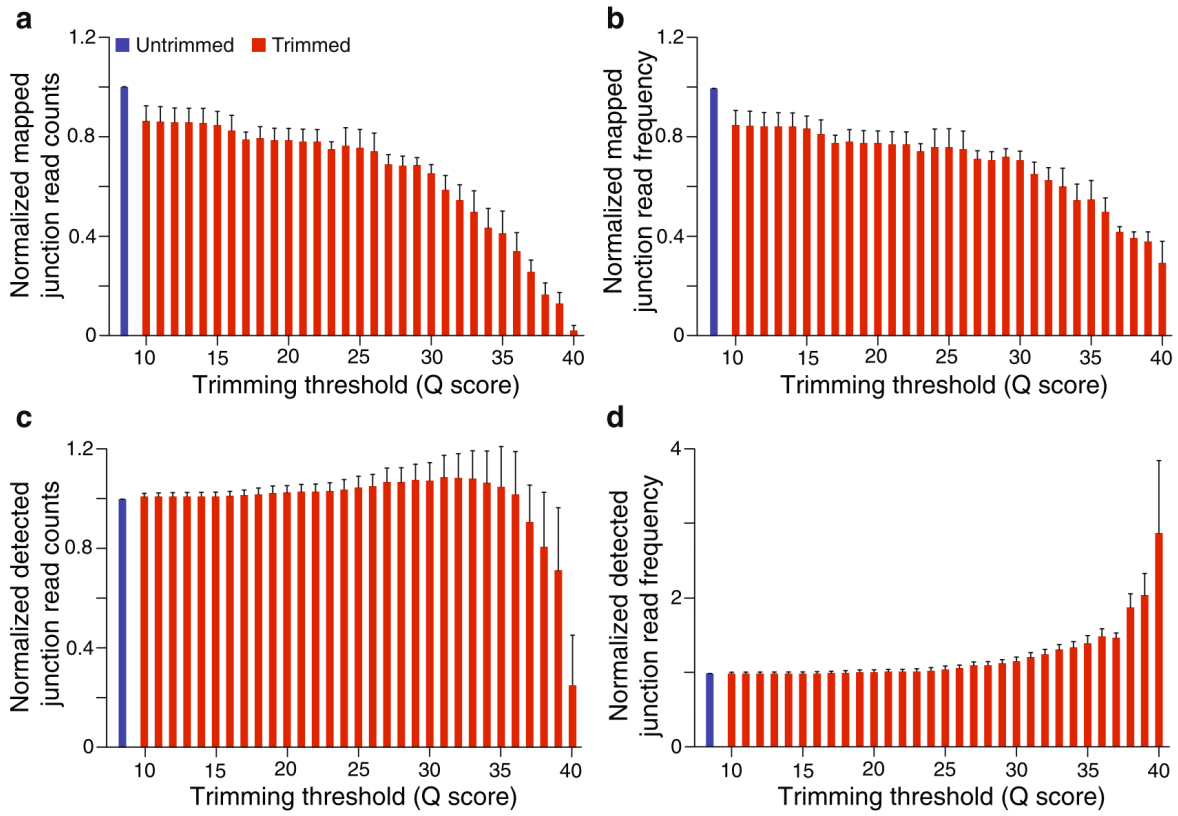

Figure 2.3 Influence of trimming on junction alignment and detection

(a) The average number of reads aligned to junctions per sample at a range of SolexaQA quality scores. (b) The average frequency of reads aligned to junctions (number of reads aligned to junctions per total reads aligned). (c) The average number of junctions detected per sample. (d) The average frequency of junction detection (number of junctions detected per total reads mapped). For all panels, data were normalized to the untrimmed value on a per sample basis. Error bars represent standard deviations.

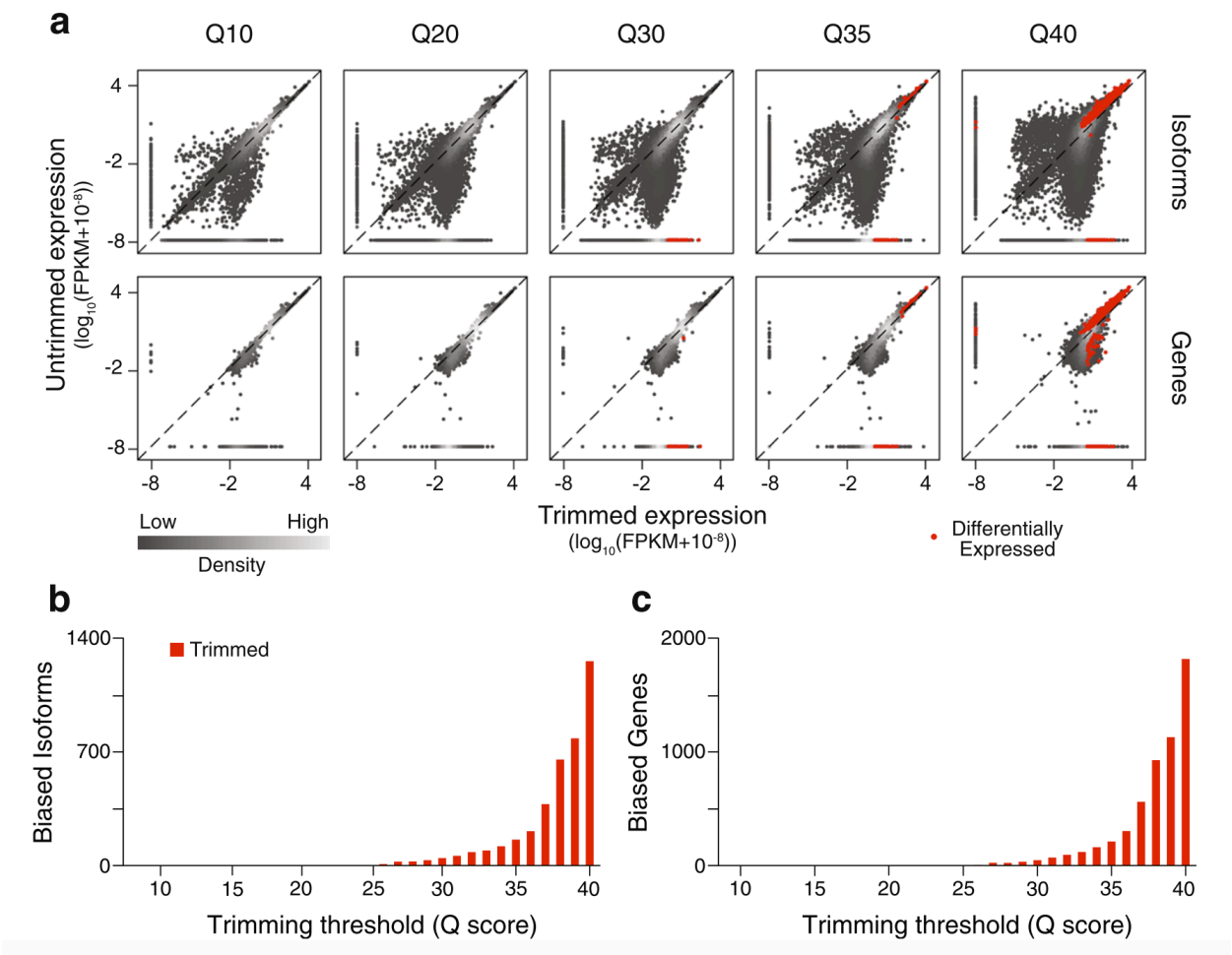

Figure 2.4 Isoform and gene expression levels after trimming

(a) Comparison of the expression estimates of isoforms and genes between increasing SolexaQA trimming thresholds (Q scores) and the untrimmed data set. Red dots represent statistically significant differential expression between data sets. (b) The number of significantly biased isoforms and (c) genes at a range of quality scores.

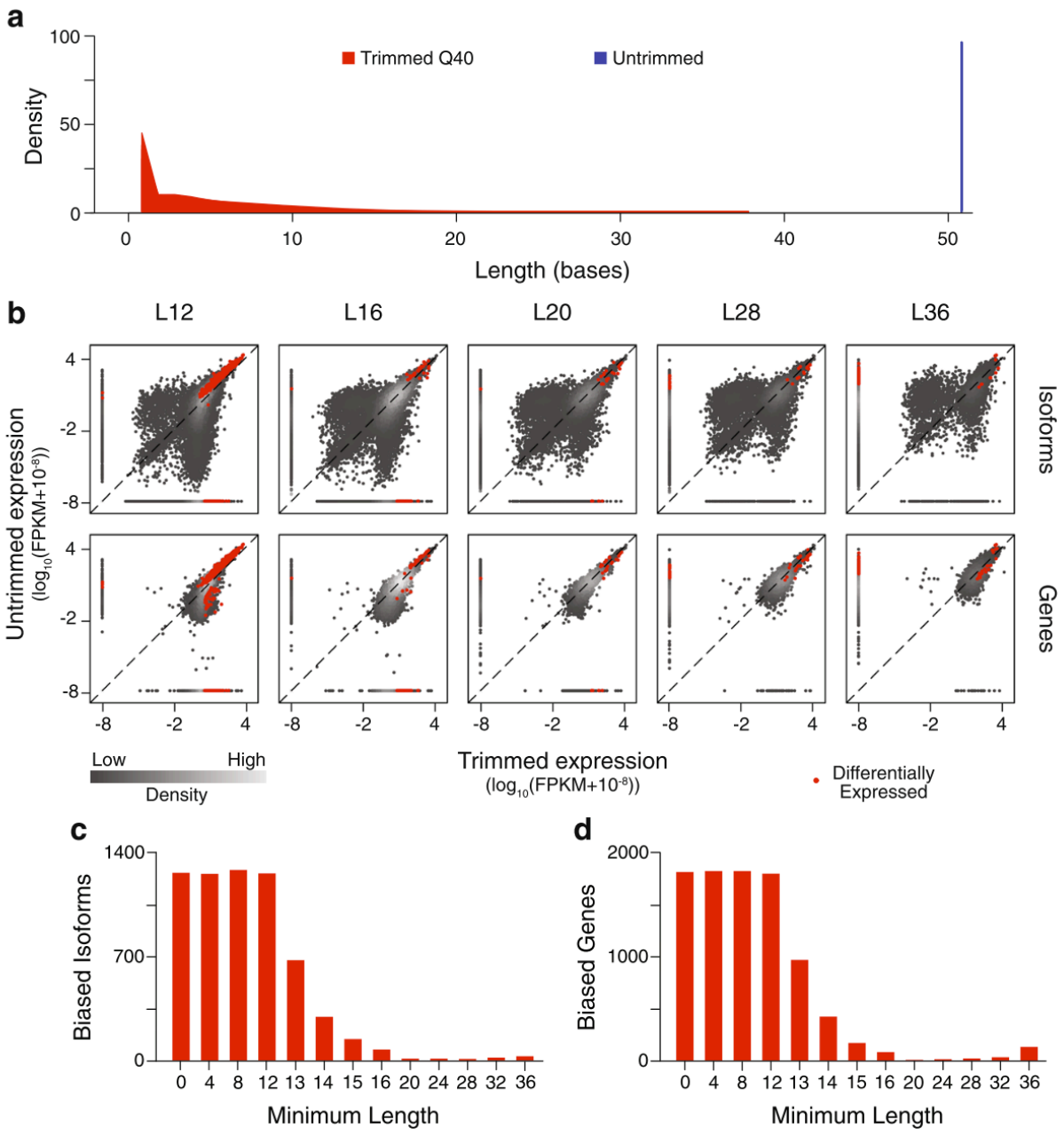

Figure 2.5 Isoform and gene expression levels after length-filtering

(a) The distribution of input read lengths before (blue) and after (red) trimming. (b) Comparison of the expression estimates of isoforms and genes between the untrimmed data set and the trimmed data sets with minimum read lengths required. L# specifies the minimum read length, in bases, required for inclusion. Red dots represent statistically significant differential expression between data sets. (c) The number of significantly biased isoforms and (d) genes at a range of minimum read lengths.

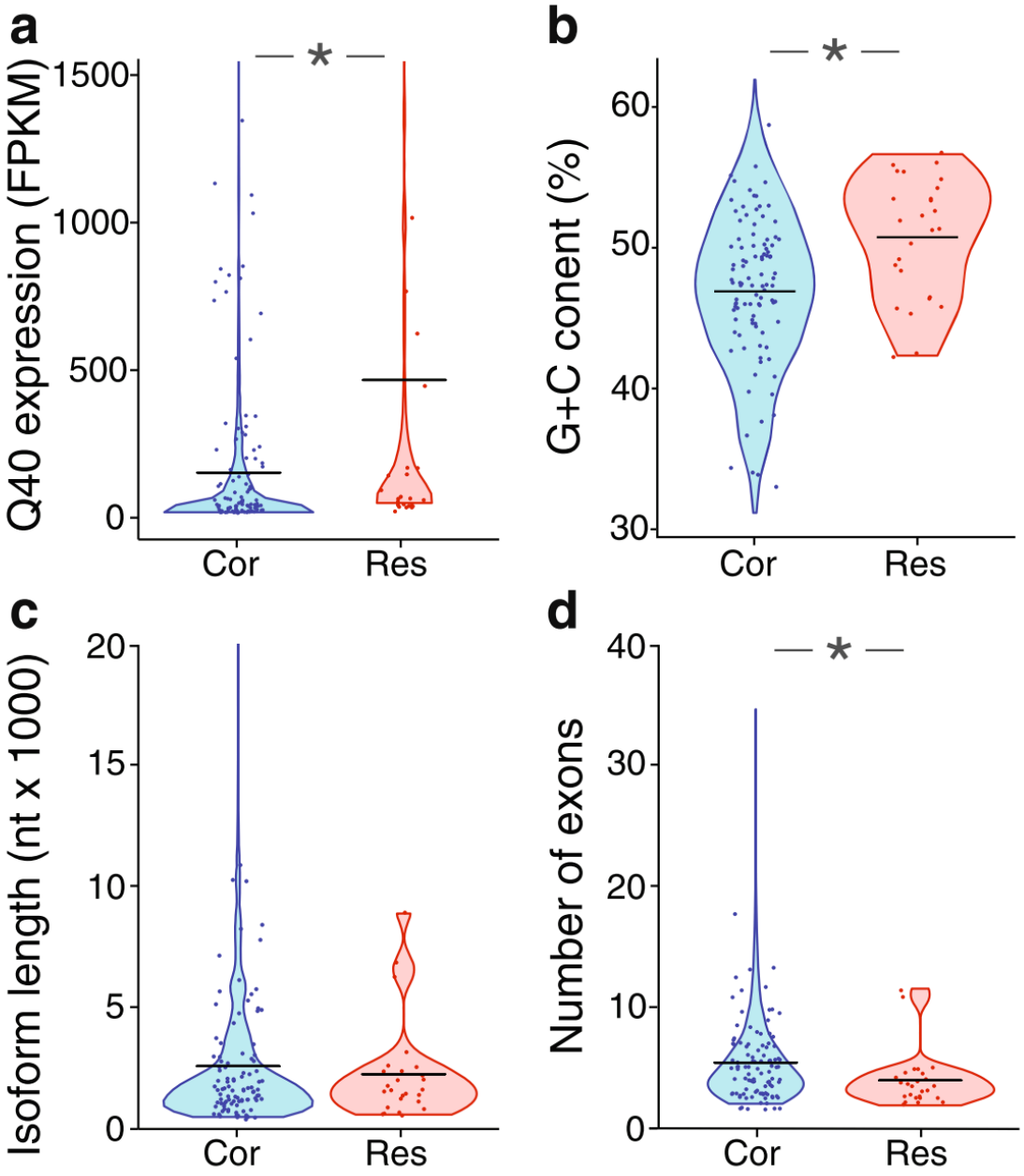

Figure 2.6 Correlations of gene and isoform properties with length filtering-resistant bias

(a) The distribution of gene expression levels (FPKM) for length filtering-correctable and -resistant genes, after SolexaQA trimming at Q=40 and prior to length filtering. (b) The fractional GC content of correctable and resistant isoforms. (c) The distribution of transcript lengths for correctable and resistant isoforms. (d) The distribution of the number of exons per isoform for correctable and resistant isoforms. *, $p<0.05$ following Benjamini-Hochberg adjustment. Bars represent the mean. For clarity, not all data points are depicted. Cor, correctable. Res, resistant.

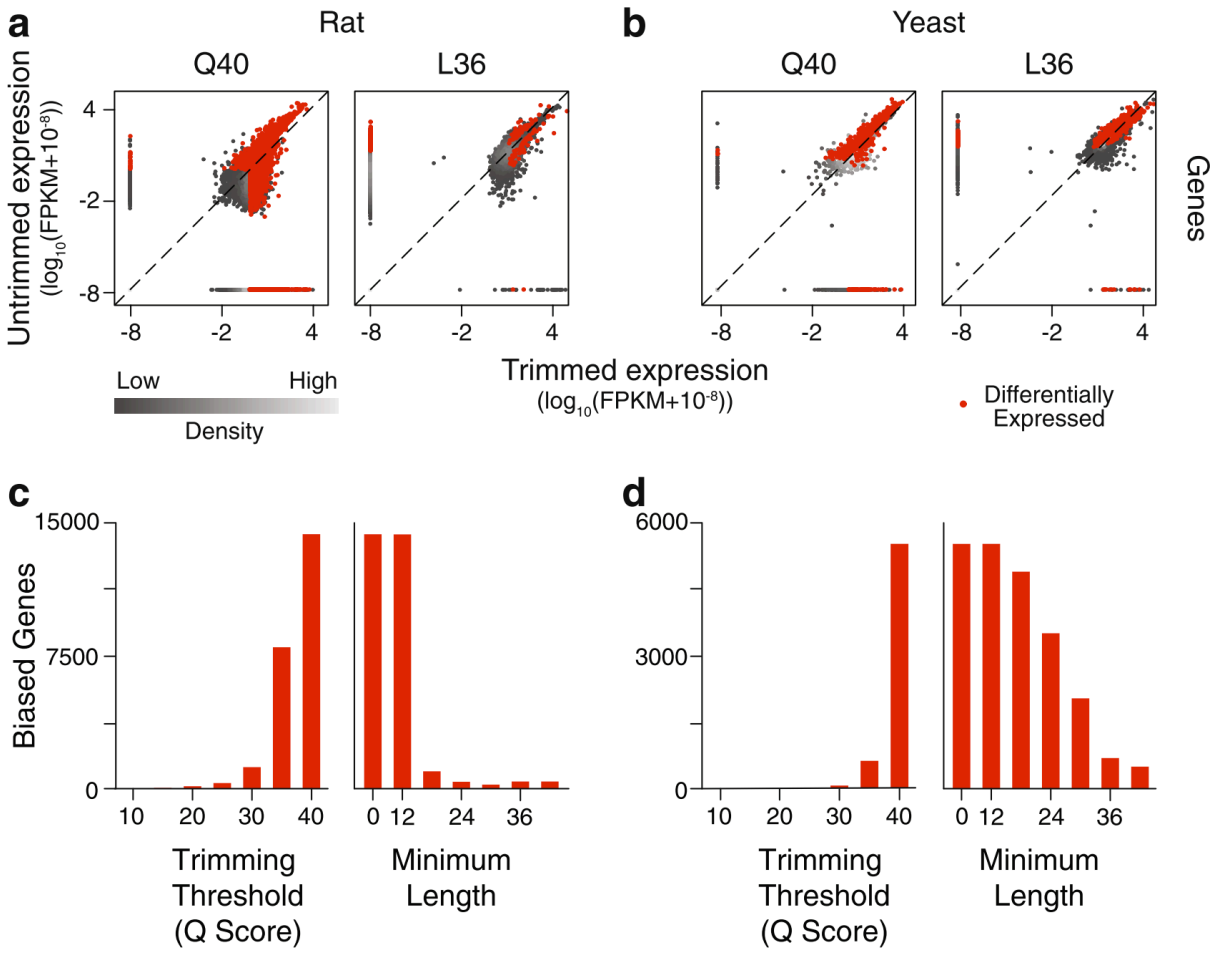

Figure 2.7 Differential gene expression following trimming in two additional data sets

(a) Comparison of the gene expression estimates from an RNA-Seq data set derived from rat samples, between SolexaQA trimmed data sets and the untrimmed data set. Q# specifies the quality threshold imposed. L# specifies the minimum read length, in bases, required for inclusion after Q40 trimming. Red dots represent statistically significant differential expression between data sets. (b) Comparison of the gene expression estimates from an RNA-Seq data set derived from yeast samples. (c) The number of significantly biased genes at a range of quality scores and lengths, rat data sets. (d) The number of significantly biased genes at a range of quality scores and lengths, yeast data sets.

2.10 Chapter Two Tables

Table 2.1 Differentially expressed genes detected by multiple analysis pipelines

Mapping Tool	DE Tool	DE Genes Q40	DE Genes Q40 L36
TopHat2	Cuffdiff2	1829	150
TopHat2 / HTSeq	DESeq	289	2
STAR	DESeq	812	2
RSEM / STAR	DESeq	79	53
STAR	EdgeR	321	0

The number of significantly differentially expressed genes detected, using 5 different analysis pipelines, when comparing the untrimmed data set to the same data set trimmed with SolexaQA, using a quality score of 40 (Q40), or with a quality score of 40 and a minimum length requirement of 36 bases (Q40 L36). DE Tool, differential expression tool. DE Genes, differentially expressed genes.

Table 2.2 Correlations between RNA-Seq gene expression estimates and microarray intensities

Data set	Untrimmed	Q10	Q20	Q30	Q40	Q40 L36
Rat	0.855	0.853	0.851	0.848	0.744	0.751
Yeast	0.891	0.891	0.889	0.887	0.863	0.785

Values represent correlation coefficients between gene expression values determined by microarray data sets and RNA-Seq data sets that were trimmed with SolexaQA with quality scores as indicated, followed by mapping and modeling with TopHat2 and Cuffdiff2.

Chapter 3 : EMPIRICAL ASSESSMENT OF ANALYSIS WORKFLOWS FOR DIFFERENTIAL EXPRESSION ANALYSIS OF HUMAN SAMPLES USING RNA-SEQ

Chapter 3 is adapted with minimal modification from:

Williams CR*, Baccarella A*, Parrish JZ, Kim CC. (2017) Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics 18(1):38.

* Authors contributed equally

3.1 Abstract

3.1a Background

RNA-Seq has supplanted microarrays as the preferred method of transcriptome-wide identification of differentially expressed genes. However, RNA-Seq analysis is still rapidly evolving, with a large number of tools available for each of the three major processing steps: read alignment, expression modeling, and identification of differentially expressed genes. Although some studies have benchmarked these tools against gold standard gene expression sets, few have evaluated their performance in concert with one another. Additionally, there is a general lack of testing of such tools on real-world, physiologically relevant datasets, which often possess qualities not reflected in tightly controlled reference RNA samples or synthetic datasets.

3.1b Results

Here, we evaluate 219 combinatorial implementations of the most commonly used analysis tools for their impact on differential gene expression analysis by RNA-Seq. A test dataset was generated using highly purified human classical and nonclassical monocyte subsets from a clinical cohort, allowing us to evaluate the performance of 495 unique workflows, when accounting for differences in expression units

and gene- versus transcript-level estimation. We find that the choice of methodologies leads to wide variation in the number of genes called significant, as well as in performance as gauged by precision and recall, calculated by comparing our RNA-Seq results to those from four previously published microarray and BeadChip analyses of the same cell populations. The method of differential gene expression identification exhibited the strongest impact on performance, with smaller impacts from the choice of read aligner and expression modeler. Many workflows were found to exhibit similar overall performance, but with differences in their calibration, with some biased toward higher precision and others toward higher recall.

3.1c Conclusions

There is significant heterogeneity in the performance of RNA-Seq workflows to identify differentially expressed genes. Among the higher performing workflows, different workflows exhibit a precision/recall tradeoff, and the ultimate choice of workflow should take into consideration how the results will be used in subsequent applications. Our analyses highlight the performance characteristics of these workflows, and the data generated in this study could also serve as a useful resource for future development of software for RNA-Seq analysis.

3.2 Background

RNA sequencing (RNA-Seq) has become the preferred technique for transcriptome-wide analysis of gene expression. However, estimating expression from short sequence reads poses unique problems such as accurate read alignment in the presence of sequencing errors, measurement bias depending on library preparation methodology, and complexity in estimating the expression of distinct mRNA transcripts with shared exons. As a result, RNA-Seq analysis is still rapidly evolving, with a wide number of tools available for each of the major processing steps, and many combinations in which these tools are commonly implemented. As such, the optimal workflow for a given application remains a subject of intensive investigation.

The most typical application of RNA-Seq is the identification of differentially expressed genes. In such an analysis, two or more conditions are compared to identify changing gene expression signatures, from which functional changes or markers of a given cellular state are inferred. The three major steps of differential expression analysis by RNA-Seq are alignment of reads to an annotated genome (or less commonly, *ab initio* reconstruction of a transcriptome annotation [1,2]), expression modeling to obtain gene-level and/or transcript-level expression estimates, and statistical analysis to identify differentially expressed genes or transcripts between comparison groups [3–8]. Various studies have evaluated the performance of the available tools at each isolated step of this workflow [9–18]; however, only a handful of studies have evaluated the performance of these approaches in concert with one another [3,19,20]. This is important since upstream processing could have substantial effects on downstream steps and outcomes [21]. In addition, performance has largely been evaluated using controlled datasets, such as those from highly purified reference RNA samples, cell lines, or reads synthetically derived *in silico*. These datasets often exhibit extreme differences in gene expression between sample groups that are unrepresentative of more typical experimental designs in which the control and test samples are more closely related to one another. In addition, such datasets do not possess the inter-sample variability in sequencing depth and quality that often occurs in many real-world settings. This is particularly true when clinical samples are involved, for which there is typically more variability in the initial sample quality, and for which analysis must also tolerate genetic variation. Thus, although such comparisons are valuable for initial benchmarking of a given algorithmic approach and its implementation, the ultimate evaluation of any given tool must take into consideration the samples to which it will be applied and the workflow context in which it will be employed.

One of the barriers to validating analysis workflows is a paucity of real-world RNA-Seq samples for which reference datasets are available for comparison. Here, we describe an RNA-Seq dataset generated from human classical and nonclassical monocyte subsets isolated to high purity. Differential gene expression analysis between these subsets has been analyzed in multiple transcriptome-wide microarray and BeadChip studies [22–25], providing us with gene sets that have been validated by multiple independent laboratories using multiple gene expression analysis platforms. Therefore, these gene sets provide a reference estimate of biological ‘truth’. Using the sequence reads from our monocyte

subset dataset, we evaluated commonly used differential expression workflows for their performance, as assessed by their agreement with these references. We find that different RNA-Seq analysis workflows differ widely in their performance, as assessed by recall, or the proportion of reference-identified genes that were also identified by the given workflow, and precision, or the proportion of genes identified by the workflow that were also identified by the reference. Many workflows perform equally well, but are calibrated differently with respect to favoring higher recall or precision, with an inverse relationship between these parameters. Based on our observations, we recommend that the selection of a given approach be guided by the tolerance of downstream applications for type I and type II errors. Used in conjunction with the previous microarray and BeadChip studies, these RNA-Seq data provide a real-world test set for guiding the development of improved software and workflows.

3.3 Methods

Samples. Blood was collected from Ugandan children as part of the *Program for Resistance, Immunology, Surveillance & Modeling of Malaria in Uganda* study using previously described methods [26]. Peripheral blood mononuclear cells (PBMCs) from a total of 18 individuals were isolated on Ficoll gradients, counted, and immediately cryopreserved and stored long-term in liquid nitrogen. Samples were thawed in the presence of DNase and immediately stained in FACS buffer with antibodies specific for the following targets: CD7 (clone 4H9), HLA-DR (clone L243), CD16 (clone CB16), CD14 (clone 61D3), CD19 (clone HIB19) from eBioscience; and CD177 (clone MEM-166) from Biolegend. For flow cytometry, classical monocytes were identified as $CD177^{-}CD7^{-}CD19^{-}HLA^{-}DR^{+}CD14^{hi}CD16^{-}$; nonclassical monocytes were identified as $CD177^{-}CD7^{-}CD19^{-}HLA^{-}DR^{+}CD14^{lo}CD16^{+}$. Both monocyte subsets were isolated to high purity using two consecutive rounds of sorting on a FACSAria, using an event rate no higher than 5,000 events/second and sorting directly into an RNA preservative buffer on the second sort. A total of 67 to 3149 cells were sorted per sample. Each sample represents a single individual, and both nonclassical and classical subsets were sorted from each individual. Sorted cells were immediately snap frozen on dry ice and stored in a -80°C freezer until the time of RNA isolation.

RNA sequencing. Cryopreserved sorted cells were thawed, and RNA was isolated using an RNAqueous Micro kit (ThermoFisher, Waltham, MA) following manufacturer recommendations with the following modifications: lysis buffer/cell aliquots were initially mixed with 180 µL of 200 proof RNase-free ethanol; the flowthrough was reloaded onto the column to capture additional material with a second binding step; and the purified RNA was eluted twice with 6 µL 55^{o} C RNase-free water following a 2 min incubation. Isolated total RNA was vacuum concentrated to 1 µL and converted to pre-amplified cDNA libraries using template-switching reverse transcription [27,28] as implemented in the SMARTer Ultra-low input kit (Clontech, Mountain View, CA). Two samples failed to yield cDNA and were thus excluded from further processing. Fragmentation was performed enzymatically using a Nextera XT DNA kit (Illumina, San Diego, CA), and barcoded samples were multiplexed, pooled, and purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA). Libraries were quality-controlled for size distribution and yield using a Bioanalyzer 2100 with high sensitivity dsDNA assay (Agilent Technologies, Santa Clara, CA), and sequenced as 51 bp single-end reads on 4 lanes of a HiSeq 2500 (Illumina) running in high-output mode at the UCSF Center for Advanced Technology (San Francisco, CA). Reads were demultiplexed with CASAVA (Illumina), and read quality assessed using FastQC [29].

Read alignment, expression modeling, and differential expression identification.

Reads were aligned to release GRCh37 of the human genome. Reads were aligned with Bowtie2, HISAT2, Kallisto, Salmon, Sailfish, SeqMap, STAR and TopHat2 [30–38]. Gene and transcript expression was estimated with BitSeq, cufflinks, htseq, IsoEM, Kallisto, RSEM, rSeq, Sailfish, Salmon, STAR, Stringtie and eXpress [32–35,37,39–45]. The IsoEM code was modified to increase the maximum available memory. Expression matrices for differential expression input were generated using custom scripts as well as the prepDE.py script provided at the Stringtie website. Differentially expressed genes or transcripts were identified with Ballgown, baySeq, BitSeq, cuffdiff, DESeq2, EBseq, edgeR exact test, limma coupled with vst or voom transformation, NBPseq, NOISeqBIO, SAMseq and Sleuth [33,39,40,46–54]. Of these, all but Ballgown, BitSeq, NBPSeq, SAMSeq, and Sleuth used intrinsic filtering or recommended extrinsic filtering of genes or transcripts prior to testing. For Sailfish and Salmon, outputs were converted to a Sleuth-ready format using wasabi [55]. For Kallisto, Sailfish, Salmon, and BitSeq,

transcript-level values were condensed to gene-level values using tximport prior to evaluating gene-level differential expression [56]. For all differential expression analyses performed at the transcript-level, significant transcripts were converted to the corresponding gene for performance evaluation, such that if a single transcript was called as differentially expressed, the corresponding gene was also called differentially expressed. We note that because of this unavoidable difference between gene-level and transcript-level comparisons, quantitative comparisons of recall and/or precision between a gene-level and a transcript-level workflow should be avoided. Rather, we recommend evaluating the relative performance of a given workflow as compared with other workflows with matched gene-level or transcript-level estimation. When possible, differential expression was assessed using multiple expression units (counts, FPKM, TPM) and performance metrics are reported separately for each unit. In general, all software was run with default parameters; specific runtime parameters are listed in Additional file 1, along with software versions, and scripts for running all code are available at https://github.com/cckim47/kimlab. Further information about implementation is available upon request. All software was run at a detection level of alpha of 0.05, FDR of 0.05, or PPLR in the most extreme 0.05. Abbreviations used throughout the figures are a six-letter code represented as AaBbCc, where Aa denotes the read aligner (RA), Bb denotes the expression modeler (EM), and Cc denotes the differential expression (DE) analysis tool. All tools and codes are shown in Table 3.1.

Preparation of reference datasets. Reference datasets were prepared from four published studies conducted on microarray or BeadChip platforms (GSE25913, GSE18565, GSE35457, GSE34515) [22–25]. An additional reference set (GSE16836 [57]) was considered, but excluded due to inter-sample variation precluding identification of differentially expressed genes. Significant differentially expressed genes between classical and nonclassical monocytes were identified for each dataset. In brief, series matrix files were downloaded from the NCBI Gene Expression Omnibus, $\log_2$ transformed if necessary, full-quantile normalized [50], and analyzed for statistically significant gene expression between classical and nonclassical monocytes. To reduce bias introduced by a single statistical method, we employed two approaches: Significance Analysis of Microarrays (SAM) [58] with a false discovery rate of 0.05, and limma [59,60], with a BH-adjusted p-value of 0.05. Performance of the workflows against both SAM and

limma were compared to one another and found to exhibit good reproducibility regardless of the statistical method used to generate the data (Additional files 2 and 3); as such, we chose to use the genes at the intersection of the two methods for our final reference gene sets.

Quantification of recall and precision. Because absolute recall and precision values are influenced by the repertoire of analytes that can be measured by a given platform, we first filtered each reference and RNA-Seq gene set to include only features measurable both by RNA-Seq (*i.e.*, present in the GRCh37 genome release) and by the microarray (*i.e.*, a probe targeting the feature was present on the microarray platform) within a given comparison. All gene set counts are reported based on these filtered numbers, as are all estimates of recall and precision. Recall was calculated as the number of significant genes in the intersection of the test RNA-Seq dataset with the reference dataset, divided by the number of genes identified as significant in the reference dataset. Precision was calculated as the number of significant genes in the intersection of the test RNA-Seq dataset with the reference dataset, divided by the number of genes identified as significant in the test RNA-Seq dataset.

3.4 Results and Discussion

3.4a Generation of a real-world RNA-Seq dataset for benchmarking

We sought to empirically assess performance characteristics of RNA-Seq analysis workflows applied to patient-derived clinical samples, which integrate multiple sources of variability that are not well represented in typical benchmarking datasets. We began by generating a test set of RNA-Seq profiles from purified human leukocytes. Specifically, we isolated cell populations from cryopreserved PBMCs collected as part of a study of malaria exposure in Ugandan children [26]. From these samples, we isolated $CD177^{-}CD7^{-}CD19^{-}HLA\text{-}DR^{+}CD14^{hi}CD16^{-}$ classical monocytes (also known as "inflammatory" monocytes) and $CD177^{-}CD7^{-}CD19^{-}HLA\text{-}DR^{+}CD14^{lo}CD16^{+}$ nonclassical monocytes (also known as "patrolling" monocytes) to high purity using two successive rounds of flow cytometry, which achieves >99% purity (Figure 3.1a). Total RNA was isolated and processed into RNA-Seq libraries using SMARTer cDNA synthesis and Nextera fragmentation and indexing. Individual samples were multiplexed and

sequenced as 51 bp single-end reads on an Illumina HiSeq 2500. Average base quality was relatively consistent across all samples, and although there was a statistically significant difference in average base quality between the classical and nonclassical monocyte groups, the effect size was small, with an absolute quality score difference of 0.4 between means (Figure 3.1b). Total reads were variable, ranging from 4 to 37 million reads per sample, but with no significant difference between the classical and nonclassical groups (Figure 3.1c). The absolute number of reads mapped by the read aligners likewise exhibited a wide range within each group, but without a significant difference between the groups (Figure 3.1d).

3.4b Overview of empirical testing

Several studies have previously explored gene expression differences between $CD14^{hi}CD16^{-}$ classical monocytes and $CD14^{lo}CD16^{+}$ nonclassical monocytes using microarray or BeadChip analysis [22–25]. Similar to our RNA-Seq dataset, these studies all represent monocytes from healthy donors. However, given that the data originate from labs in Singapore, the United States, and Germany, it is likely that there is some bias in genetics across the studies. It is also likely that these microarray data do not reflect the same genetic makeup and environmental pressures present in our data, which are obtained from Ugandan children with a high degree of malaria exposure. It should also be noted that recent studies have differentiated between three, rather than two, monocyte subsets [61], and several reference datasets were produced prior to this advancement and thus might not represent the same degree of purity in their nonclassical monocyte subset [22,24,25]. Despite these differences, in aggregate, these datasets provide a strong reference of biological 'truth' for comparison, as individual datasets can be evaluated as independent assessments of a given RNA-Seq analysis workflow. Because differentially expressed gene lists were not available for all studies and statistical criteria differed between studies, we have made our re-analysis of these publicly available datasets available as supplementary data (Additional file 2). Overall, the four datasets identified 4069 unique genes. Of these, 572 were shared among all 4 datasets, and 2755 were shared between at least two datasets. The Wong dataset showed the least overlap with the other datasets, contributing approximately half of the genes unique to a single dataset (Figure 3.2).

With these four datasets as our references for performance comparisons, we focused our evaluation on RNA-Seq analysis approaches that have gained wide adoption due to their performance, availability, documentation, and/or ease of implementation. We evaluated 9 read aligners, 12 expression modelers and 13 methods for identifying differentially expressed genes and transcripts (Table 3.1), in all possible combinations. Exceptions included cases in which the output of an earlier stage was incompatible as the input to a later stage due to file format or expression units, or difficulty with software execution. In total, including comparisons made at the gene level and transcript level, and comparisons using expression data reported in counts, TPMs, or FPKMs, we evaluated 495 unique workflows (Additional File 4). We note that some of the workflows were not intended to be used in the resulting combinations by the original authors of the software.

Despite the aforementioned heterogeneity in the microarray and BeadChip analysis results, we found that performance of various RNA-Seq workflows was remarkably consistent across all four reference datasets. We note, however, that these reference datasets are also subject to the inherent biases of the experimental and computational methods used to produce them. Here, we have depicted our results using performance metrics averaged across all four references; however, we have also made available the performance estimates for each individual reference (Additional file 5 and Additional file 6), and an interactive visualization to explore the relative performance of the tools in more detail (Additional file 7).

3.4c Differential influence of workflow stages

For each workflow consisting of all three steps (read alignment, expression modeling, and identification of differentially expressed genes), we evaluated the ability to detect genes differentially expressed between classical and nonclassical monocytes. When workflows identified a differentially expressed transcript, the corresponding gene was annotated as significant for performance evaluations, regardless of the status of other transcripts of the gene. In general, more significant genes were observed when evaluations were performed at the transcript level, because there are more transcripts than genes to potentially be differentially expressed. We have separated the analyses performed at the gene and transcript levels to highlight this difference throughout, and recommend that direct comparisons across

these units not be made. Across workflows, we observed substantial variability in the number of differentially expressed genes identified (n=208 to 9,489 significant genes; Figure 3.3 and Additional file 5). Beyond the overall variation, two trends were apparent when the number of genes identified was examined on a by-tool basis. First, the differential expression tool had a larger impact on the number of genes identified than the read aligner and expression modeler (Figure 3.3), as demonstrated by the relative homogeneity of range, distribution, and medians of the first two steps compared to the more variable parameters for the final step. Consequently, the coefficient of variation of the medians was largest for differential expression tools, as compared to read aligners and expression modelers, when assessed at both the gene level (20.5 versus 9.9 and 9.8, respectively) and the transcript level (43.4 versus 10.8 and 39.3). Second, differential expression tools varied in their robustness to different inputs, with some tools exhibiting relatively reproducible predictions regardless of the read aligner and expression modeler choices and expression units (*e.g.*, Ballgown), and other differential expression analysis tools exhibiting a wide range of predictions as the input parameters varied (*e.g.*, NOISeqBIO at the gene level) (Figure 3.3e, f).

We also evaluated performance of the workflows by calculating recall (intersecting significant genes divided by total number of significant reference genes) and precision (intersecting significant genes divided by total number of significant genes identified by RNA-Seq), using the microarray datasets as references. In order to further examine the influence of each stage of the workflow on the prediction of differentially expressed genes, we computed the absolute difference in recall and precision in all possible pairwise comparisons of workflows differing in only one component. Similar to the impact on the number of genes identified, for both precision and recall, the largest effects were observed in workflows differing in the statistical analysis of differential expression, as indicated by the increased medians of differences for this step (Figure 3.4).

3.4d Heterogeneity in performance characteristics of different workflows

We next evaluated performance by examining the specific recall and precision for individual workflows. Recall across the workflows was highly correlated with the number of genes identified (Figure 3.5a, b). This was true regardless of which of the reference datasets was used for comparison (Additional

file 5 and Additional file 6). Furthermore, the relative rankings of the workflows, ordered by absolute recall value, tended to be consistent across reference datasets (Additional file 6). For gene-level predictions, a subset of workflows using SAMseq exhibited the highest recall values; for transcript-level predictions, workflows using baySeq and NBPSeq exhibited the highest recall (Figure 3.5a, b). However, there were exceptions to these rules, depending on the choice of read aligner and expression modeler (Figure 3.5 and Additional file 6).

Precision was highly inversely correlated with the number of genes predicted across the workflows (Figure 3.5c, d). Like recall, rankings were generally consistent regardless of which reference dataset was used, as was the overall relationship between significant genes and precision (Additional file 5 and Additional file 6). For gene-level predictions, a subset of workflows using NOISeqBIO exhibited the highest precision, whereas for transcript-level predictions those with the highest precision used several different combinations of tools, with the most prevalent being Ballgown and NOISeqBIO. Strikingly, when used on transcript-level data, the commonly used combination of TopHat2, cufflinks and cuffdiff exhibited one of the highest precision values, coupled with the second lowest number of differentially expressed genes identified (Figure 3.5 and Additional file 5).

3.4e Performance tradeoff

It is important to note that the specific workflows highlighted above are at the extremes of one or another performance metric. As would be expected, the prediction of more or fewer significant genes results in a tradeoff between recall and precision. For example, the workflows employing NOISeqBIO that exhibit the highest precision were also among those with the lowest recall (Figure 3.5 and Additional file 6). An investigation of the relationship between precision and recall revealed that this tradeoff generally persisted throughout, with many workflows following an inverse linear relationship between precision and recall (Figure 3.6a, b). This held true for both gene- and transcript-level analysis, was true regardless of the expression estimation units, and was also consistent across reference datasets (Figure 3.6a, b, Additional file 7, and Additional file 8).

As observed previously with the number of significant genes and performance differences by step, the differential expression step had the greatest impact on the performance of each workflow along

the spectrum of recall and precision (Figure 3.6c, d). Specific tools that tended to track along this linear tradeoff were Ballgown, DESeq2, limma + voom, limma + vst and SAMseq; baySeq and EBseq consistently deviated the furthest. SAMseq, one tool with a nonparametric approach, has been highlighted as a high performer previously [3,16], in particular when there are a large number of replicates available to approximate the underlying distribution, as is the case here; it performs well, though it does exhibit a tendency toward higher recall at the expense of precision. NOISeqBIO, the other tested differential expression tool that assumes a nonparametric distribution, has previously been observed to identify fewer differentially expressed genes with larger sample sizes [3]; we also observe this, as well as correspondingly low recall values. Of the differential expression methods tested, baySeq and EBseq are the most similar to each other in underlying statistical methodology; both use an underlying negative binomial model, and then estimate a posterior probability of being differentially expressed for each gene [46,48]. The observation that EBseq deviated furthest from the precision/recall performance line, due to decreased precision without gains in recall, is similar to previous observations showing that EBSeq tended to produce many false positives with large sample sizes [16]. When applied to gene-level data, baySeq performed similarly to EBseq though not as extreme, with relatively low recall without commensurate gains in precision, which may reflect the similarity in their underlying methods. The development of Ballgown drew on the limma statistical methodologies based on linear models, although only Ballgown (and not limma) can accept TPM and FPKM data, in addition to counts. All three linear model workflows perform well and track along the linear precision/recall tradeoff, irrespective of upstream processing. However, there is some difference in default tuning, as Ballgown results tended towards higher precision, whereas limma + voom and limma + vst tended towards higher recall.

Aligners and estimators generally did not follow any specific trends, consistent with our observation that their influence is overshadowed by that of the differential expression analysis tool. However, two exceptions stood out. First, using BitSeq as the expression modeler tended to result in identification of large numbers of differentially expressed genes, but only in combination with differential expression tools that used an underlying negative binomial model for expression data (BaySeq, DESeq2, edgeR, and NBPSeq); EBSeq was the one exception, with the number of differentially expressed genes within range of workflows using differential expression tools that model other distributions (Ballgown,

BitSeq, limma, and NOISeqBIO). We note that BitSeq was unusual in that its most prevalent estimated expression count value was between 1 and 2, rather than less than 1 as most expression modelers estimated; this likely explains why these expression data were poorly modeled by a negative binomial distribution. Second, using STAR as the read aligner, most notably with Ballgown as the differential expression tool, led to some of the highest performance workflows having a balance of precision and recall. Interestingly, these best performing workflows are not combinations of aligner and estimator that are suggested by the Ballgown authors, demonstrating the utility of broad, empirical exploration for uncovering improved workflows. Overall, there are multiple workflows that exhibit excellent performance, and, the relationship between recall and precision among the differential expression workflows that track along the inverse linear relationship likely reflects differential calibration of these methods with regard to the tradeoff between sensitivity and specificity, rather than any fundamental difference in statistical or algorithmic performance.

The above observations also suggest that the selection of a specific workflow should be largely influenced by the tolerance of a specific application for type I versus type II errors. However, it is also important to note that a significant number of workflows deviated from the roughly linear relationship between recall and precision, particularly for tools targeted at gene-level analyses; such workflows could be considered to exhibit lower performance, as higher performance workflows would be available as alternatives at a given recall or precision target value. Furthermore, our findings reflect a defined set of parameters, such as read length, sequencing coverage, sample number, and genetic polymorphism. Thus, it is possible that the performance, both absolute and relative, of the above workflows could vary under other conditions, as some studies have observed [8,16]; as such, additional studies comparing workflow performance will be required to understand the generalizability of our observations. Importantly, when selecting a pipeline it is essential to consider not only the specific tools selected at each stage of the workflow, but also how they interact with one another.

3.5 Conclusions

The choice of RNA-Seq analysis workflow, applied to genotypically heterogeneous samples, exerts significant influence on the repertoire, recall, and precision of the differentially expressed gene set

that is identified. The impact of software selection at each step was not simply a function of upstream position in the workflow; rather, the choice of differential expression analysis approach exhibited the strongest impact on recall and precision, with more modest influences from the read aligner and expression modeler. The ultimate choice of workflow should take into consideration how the results will be used, and the performance characteristics described in this study. These, used in conjunction with consideration of the tolerance of the downstream applications for type I and type II errors, can guide the selection of an appropriate workflow. The data generated in this study also provide a useful benchmarking set for further development of RNA-Seq analysis tools and workflows.

3.6 Declarations

List of Abbreviations Used

RNA-Seq: RNA sequencing

PBMCs: peripheral blood mononuclear cells

Ethics approval and consent to participate

Written informed consent was obtained from the parent or guardian of all study participants. The study protocol was approved by the Uganda National Council of Science and Technology and the institutional review boards of the University of California, San Francisco (UCSF), Makerere University, and the Centers for Disease Control and Prevention.

Consent for publication

Not applicable.

Availability of data and material

The human monocyte RNA-Seq data set generated in this article is available in the NCBI Sequence Read Archive (SRA) under accession number SRP082682. The monocyte microarray data sets used were obtained from the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE25913,

GSE18565, GSE35457, and GSE34515. Scripts for running all code are available in a GitHub repository at https://github.com/cckim47/kimlab.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a grant from the National Institutes of Health, University of California, San Francisco-Gladstone Institute of Virology & Immunology Center for AIDS Research, P30 AI027763, NIAID U19 AI089674, NIAID R21 AI114916, NEI U10 EY008057, and NIDDK P30 DK063720 to CCK; a National Institutes of Health grant NINDS R01 NS076614 and a UW Research Innovation award to JZP; an ACCMA Community Health Foundation Summer Scholarship and a Schoeneman Scholarship to AB; and an NSF Graduate Research Fellowship (DGE1256032) to CRW.

Author's contributions

AB generated the samples. CRW, AB, and CCK performed analysis. CRW, AB, JZP, and CCK wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Dr. Thomas Sandmann for data visualization suggestions, Dr. Maggie Feeney for her collaboration, and Dr. Michelle Boyle for technical assistance.

3.7 Supplemental Information

Additional files referenced in the text of this chapter are available with the published version of this manuscript, found at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1457-z. The content of these files is as follows:

Additional file 1: Table of software tools, with versions and runtime parameters.

Format: XLSX Size: 16.4 KB.

This file can be viewed with: Microsoft Excel

Additional file 2: Table of genes differentially expressed between nonclassical and classical monocytes in four reference studies, identified using limma and SAM, and their intersections with each other and annotated features in GRCh37.

Format: XLSX Size: 2.65 MB.

This file can be viewed with: Microsoft Excel

Additional file 3: Figure of similarity in performance characteristics of significant gene identification by limma and SAM. Ranks of absolute precision and recall are shown for each workflow, when comparing SAM and limma microarray analysis of the reference datasets (a), comparing SAM and the intersection of SAM and limma (b), or comparing limma and the intersection of SAM and limma (c).

Format: PDF Size: 786 KB.

This file can be viewed with: Adobe Acrobat Reader

Additional file 4: Figure of all workflow and unit combinations run. (a) Gene-level workflows. (b) Transcript-level workflows.

Format: PDF Size: 718 KB.

This file can be viewed with: Adobe Acrobat Reader

Additional file 5: Table of number of significant genes identified, for each workflow against each reference dataset.

Format: XLSX Size: 46.5 KB.

This file can be viewed with: Microsoft Excel

Additional file 6: Table of workflow performance, including values and ranks for recall and precision, for each workflow against each reference dataset.

Format: XLSX Size: 189 KB.

This file can be viewed with: Microsoft Excel

Additional file 7: Interactive figure of comparison of performance metrics. (a) Absolute precision and recall for each workflow. (b) Relative ranks of precision and recall for each workflow.

Format: XLSX Size: 688 KB.

This file can be viewed with: Microsoft Excel

Additional file 8: Figure of recall and precision, for each reference dataset. Precision and recall as assessed using the Ingersoll (a,b), Haniffa (c,d), Frankenberger (e,f), and Wong (g, h) references, with top (shaded) and balanced (white) performers labeled.

Format: PDF Size: 186 KB.

This file can be viewed with: Adobe Acrobat Reader

3.8 References

1. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat. Methods. 2010;7:909–12.

2. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 2011;29:644–52.

3. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief. Bioinform. 2015;16:59–70.

4. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat. Methods. 2009;6:S22–32.

5. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. 2010;11:220.

6. Poplawski A, Marini F, Hess M, Zeller T, Mazur J, Binder H. Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective. Brief. Bioinform. 2015;

7. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods. 2011;8:469–77.

8. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16:150.

9. Fonseca NA, Marioni J, Brazma A. RNA-Seq gene profiling--a systematic empirical comparison. PloS One. 2014;9:e107026.

10. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods. 2013;10:1185–91.

11. Palmieri N, Nolte V, Suvorov A, Kosiol C, Schlötterer C. Evaluation of different reference based annotation strategies using RNA-Seq - a case study in Drososphila pseudoobscura. PloS One. 2012;7:e46415.

12. Benjamin AM, Nichols M, Burke TW, Ginsburg GS, Lucas JE. Comparing reference-based RNA-Seq mapping methods for non-human primate data. BMC Genomics. 2014;15:570.

13. Reddy R. A Comparison of Methods: Normalizing High-Throughput RNA Sequencing Data. bioRxiv. 2015;026062.

14. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am. J. Bot. 2012;99:248–56.

15. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. PloS One. 2014;9:e103207.

16. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91.

17. Tang M, Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC Bioinformatics. 2015;16:361.

18. Yang C, Wu P-Y, Tong L, Phan JH, Wang MD. The impact of RNA-seq aligners on gene expression estimation. ACM-BCB ACM Conf. Bioinforma. Comput. Biol. Biomed. ACM Conf. Bioinforma. Comput. Biol. Biomed. 2015;2015:462–71.

19. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. 2012;40:10084–97.

20. Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 2016;17:74.

21. Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. Genome Biol. 2015;16:177.

22. Ingersoll MA, Spanbroek R, Lottaz C, Gautier EL, Frankenberger M, Hoffmann R, et al. Comparison of gene expression profiles between human and mouse monocyte subsets. Blood. 2010;115:e10–9.

23. Wong KL, Tai JJ-Y, Wong W-C, Han H, Sem X, Yeap W-H, et al. Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. Blood. 2011;118:e16–31.

24. Haniffa M, Shin A, Bigley V, McGovern N, Teo P, See P, et al. Human tissues contain CD141hi cross-presenting dendritic cells with functional homology to mouse CD103+ nonlymphoid dendritic cells. Immunity. 2012;37:60–73.

25. Frankenberger M, Hofer TPJ, Marei A, Dayyani F, Schewe S, Strasser C, et al. Transcript profiling of CD16-positive monocytes reveals a unique molecular fingerprint. Eur. J. Immunol. 2012;42:957–74.

26. Kamya MR, Arinaitwe E, Wanzira H, Katureebe A, Barusya C, Kigozi SP, et al. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. Am. J. Trop. Med. Hyg. 2015;92:903–12.

27. Matz M, Shagin D, Bogdanova E, Britanova O, Lukyanov S, Diatchenko L, et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Res. 1999;27:1558–60.

28. Petalidis L, Bhattacharyya S, Morris GA, Collins VP, Freeman TC, Lyons PA. Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. Nucleic Acids Res. 2003;31:e142.

29. Babraham Bioinformatics. FastQC at Babraham Bioinformatics [Internet]. Babraham Bioinforma. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012;9:357–9.

31. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods. 2015;12:357–60.

32. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 2016;34:525–7.

33. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 2016;11:1650–67.

34. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat. Biotechnol. 2014;32:462–4.

35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 2013;29:15–21.

36. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.

37. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. bioRxiv. 2016;021592.

38. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinforma. Oxf. Engl. 2008;24:2395–6.

39. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 2013;31:46–53.

40. Glaus P, Honkela A, Rattray M. Identifying differentially expressed transcripts from RNA-seq data with biological variation. Bioinforma. Oxf. Engl. 2012;28:1721–8.

41. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinforma. Oxf. Engl. 2015;31:166–9.

42. Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms Mol. Biol. AMB. 2011;6:9.

43. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

44. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinforma. Oxf. Engl. 2009;25:1026–32.

45. Roberts A, Pachter L. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods. 2013;10:71–3.

46. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422.

47. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

48. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinforma. Oxf. Engl. 2013;29:1035–43.

49. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. 2010;26:139–40.

50. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 2004;3:Article3.

51. Yanming D, W SD, S CJ, H CJ. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Stat. Appl. Genet. Mol. Biol. 2011;10:1–28.

52. Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43:e140.

53. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res. 2013;22:519–36.

54. Pimentel HJ, Bray N, Puente S, Melsted P, Pachter L. Differential analysis of RNA-Seq incorporating quantification uncertainty. bioRxiv. 2016;058164.

55. wasabi [Internet]. GitHub. [cited 2016 Aug 31]. Available from: https://github.com/COMBINE-lab/wasabi

56. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4:1521.

57. Ancuta P, Liu K-Y, Misra V, Wacleche VS, Gosselin A, Zhou X, et al. Transcriptional profiling reveals developmental relationship and distinct biological functions of CD16+ and CD16- monocyte subsets. BMC Genomics. 2009;10:403.

58. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U. S. A. 2001;98:5116–21.

59. Kim CC, Falkow S. Significance analysis of lexical bias in microarray data. BMC Bioinformatics. 2003;4:12.

60. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, editors. Bioinforma. Comput. Biol. Solut. Using R Bioconductor [Internet]. New York, NY: Springer New York; 2005. p. 397–420. Available from: http://dx.doi.org/10.1007/0-387-29362-0_23

61. Wong KL, Yeap WH, Tai JJY, Ong SM, Dang TM, Wong SC. The three human monocyte subsets: implications for health and disease. Immunol. Res. 2012;53:41–57.

3.9 Chapter Three Figures

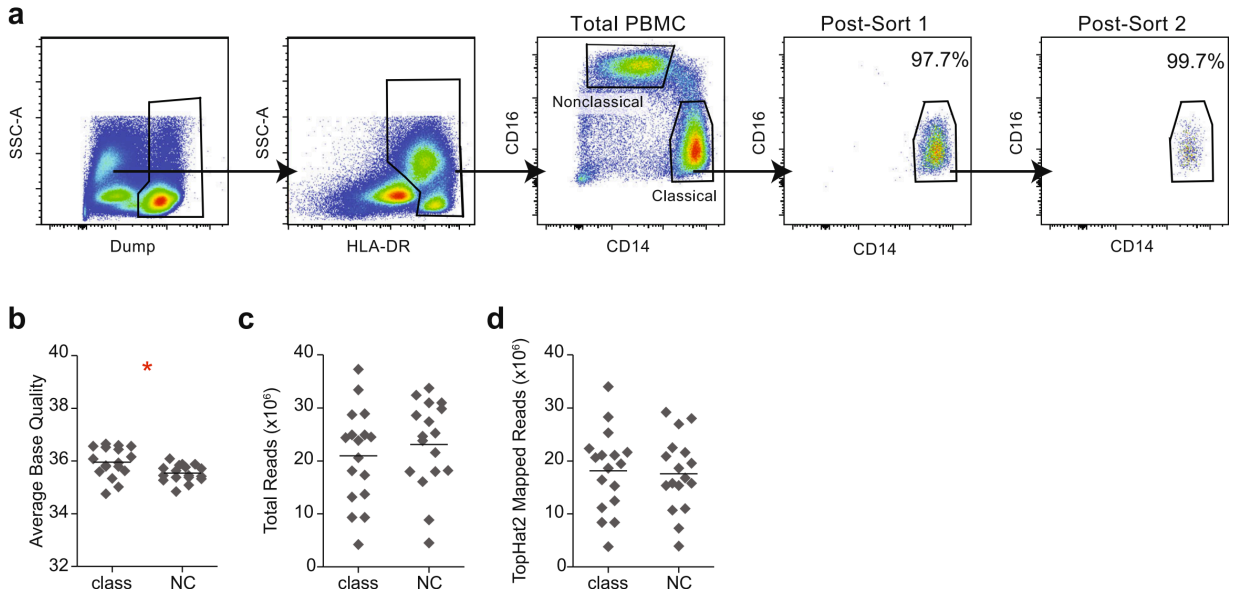

Figure 3.1 Monocyte isolation by flow cytometry and sequence read characteristics

(a) Gating strategy for isolation of monocyte subsets, and a representative demonstration of increasing purity of monocyte subsets upon successive rounds of flow cytometric sorting. (b) Average base quality across all bases within a sample. (c) Total reads per sample. (d) Representative example of total reads mapped to the human genome. Class, classical monocytes; NC, nonclassical monocytes.

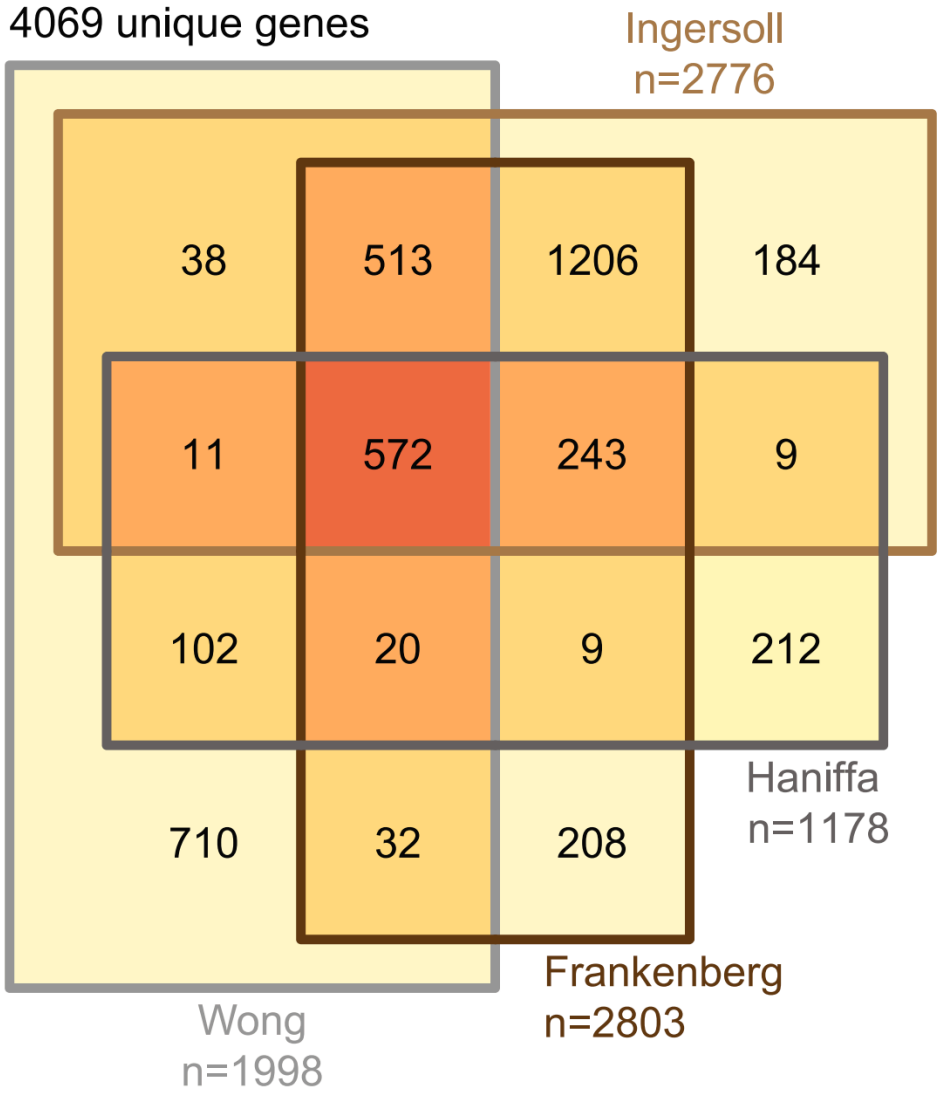

Figure 3.2 Concordance between significant gene expression differences between classical and nonclassical monocytes identified in four independent studies

Venn diagram showing degree of overlap of genes identified as significant by both SAM and limma from each microarray or BeadChip study.

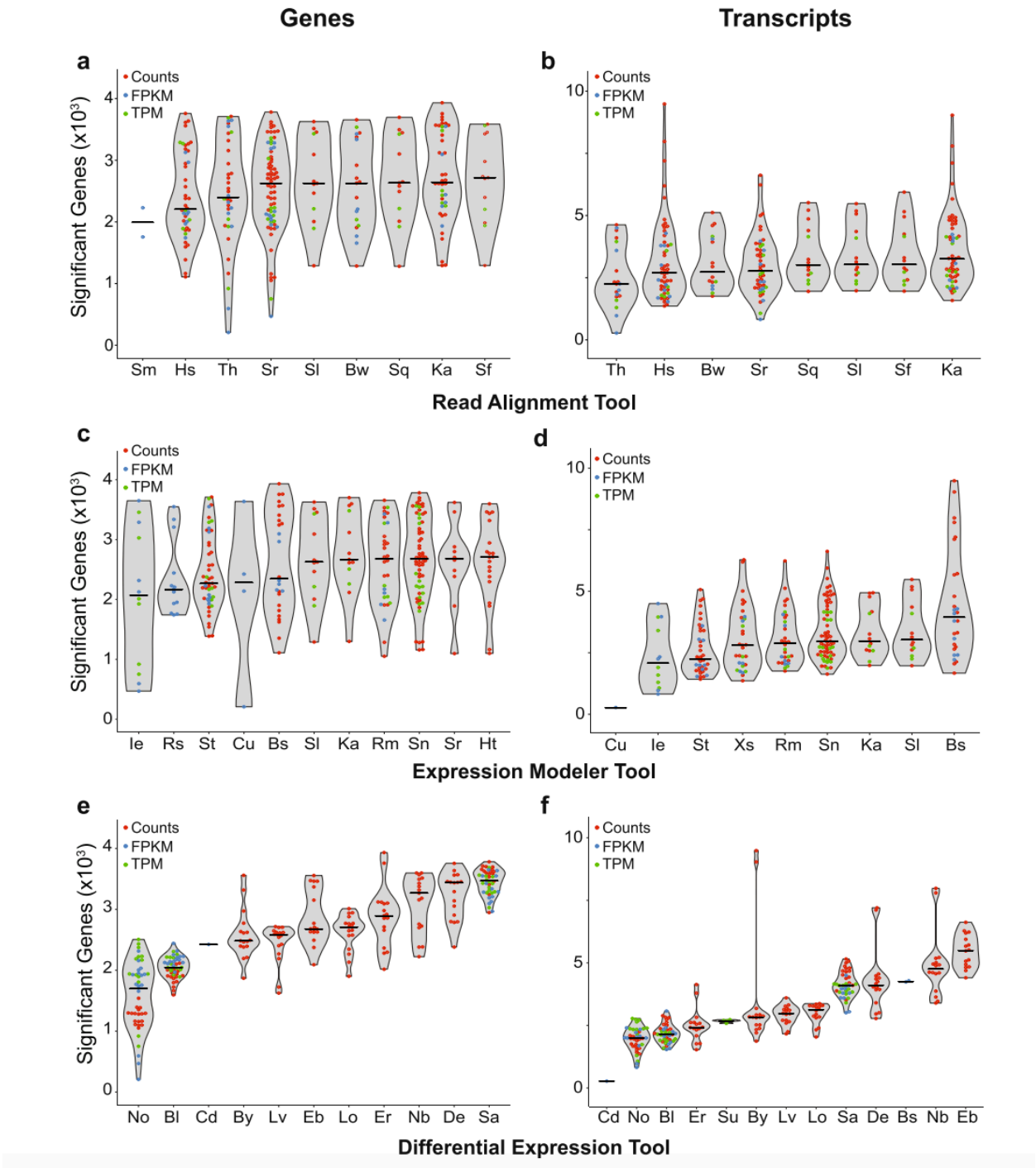

Figure 3.3 Number of significant genes predicted by workflows using a given method

The number of genes predicted by each workflow using a given read aligner (a, b), expression modeler (c, d), or differential expression tool (e, f), split by analyses run at the gene (a, c, e) or transcript (b, d, f) level. Each point represents a single workflow; line shows median.

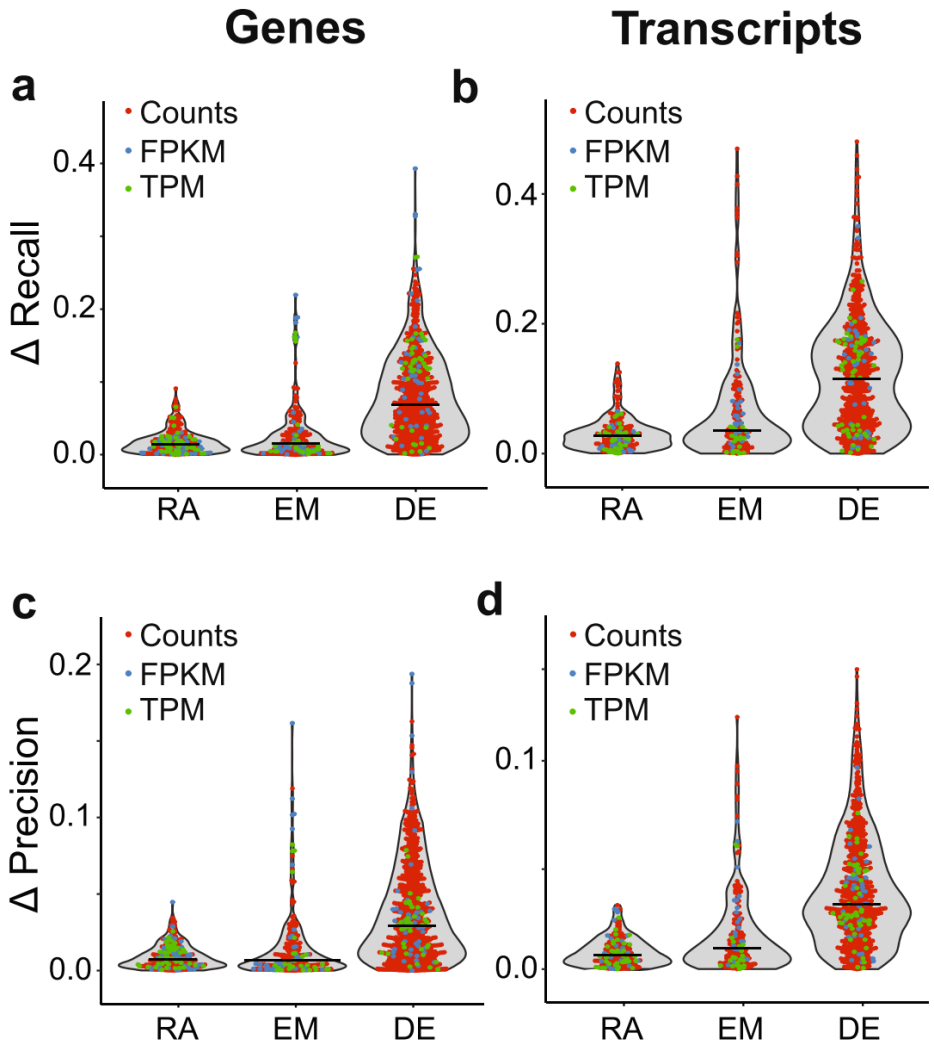

Figure 3.4 Impact of individual stages of the workflow on overall performance characteristics

The difference in recall and precision was calculated for exhaustive pairwise comparisons of workflows in which the software used for the given stage under evaluation was varied while the two other tools were held constant. The points reflect each absolute difference; the line represents the median. RA, read aligner; EM, expression modeler; DE, differential expression.

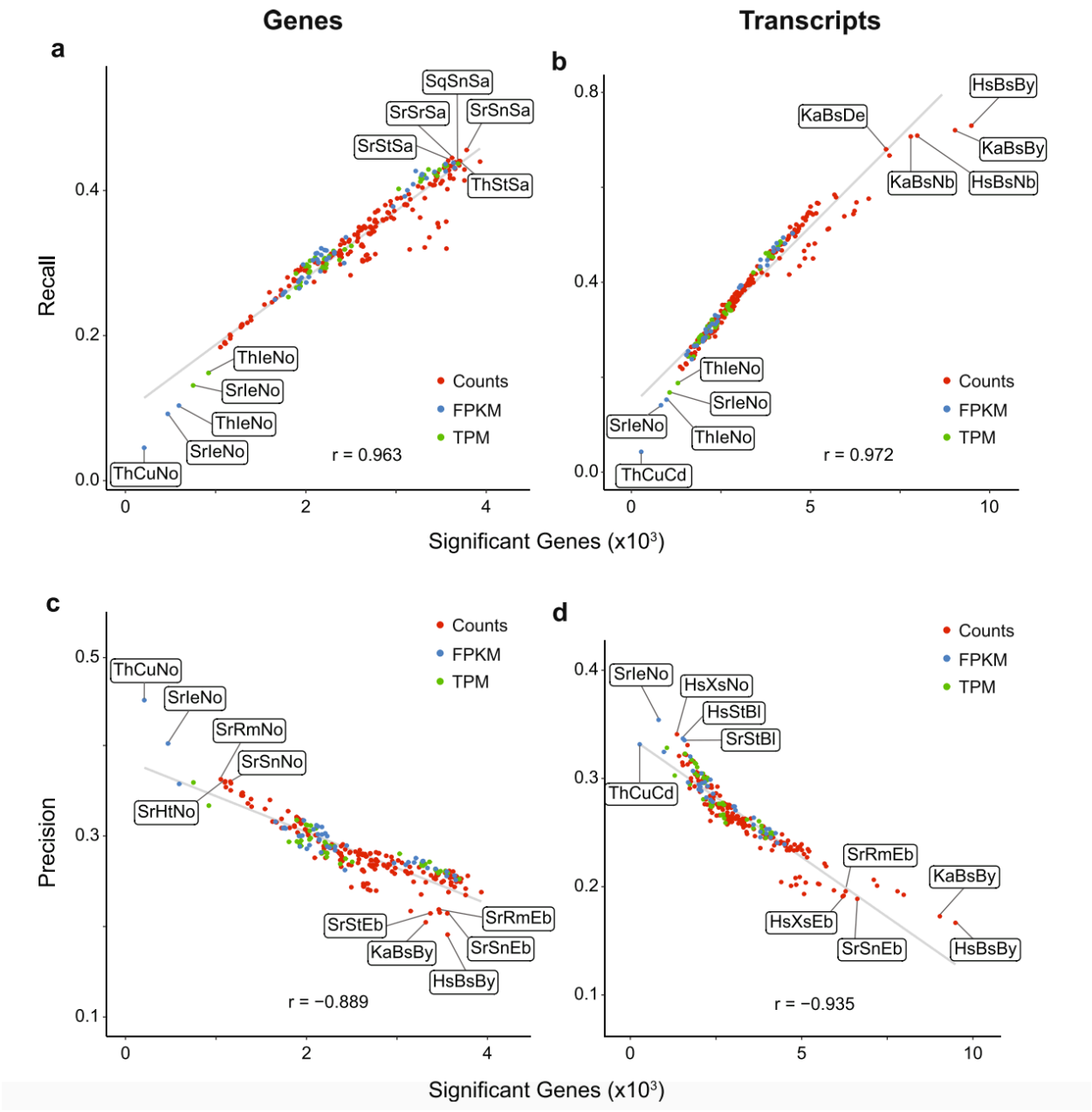

Figure 3.5 Relationship of recall and precision with number of genes predicted

Recall (a, b) and precision (c, d), plotted vs number of significant genes predicted by each workflow. Pearson *r* values are shown.

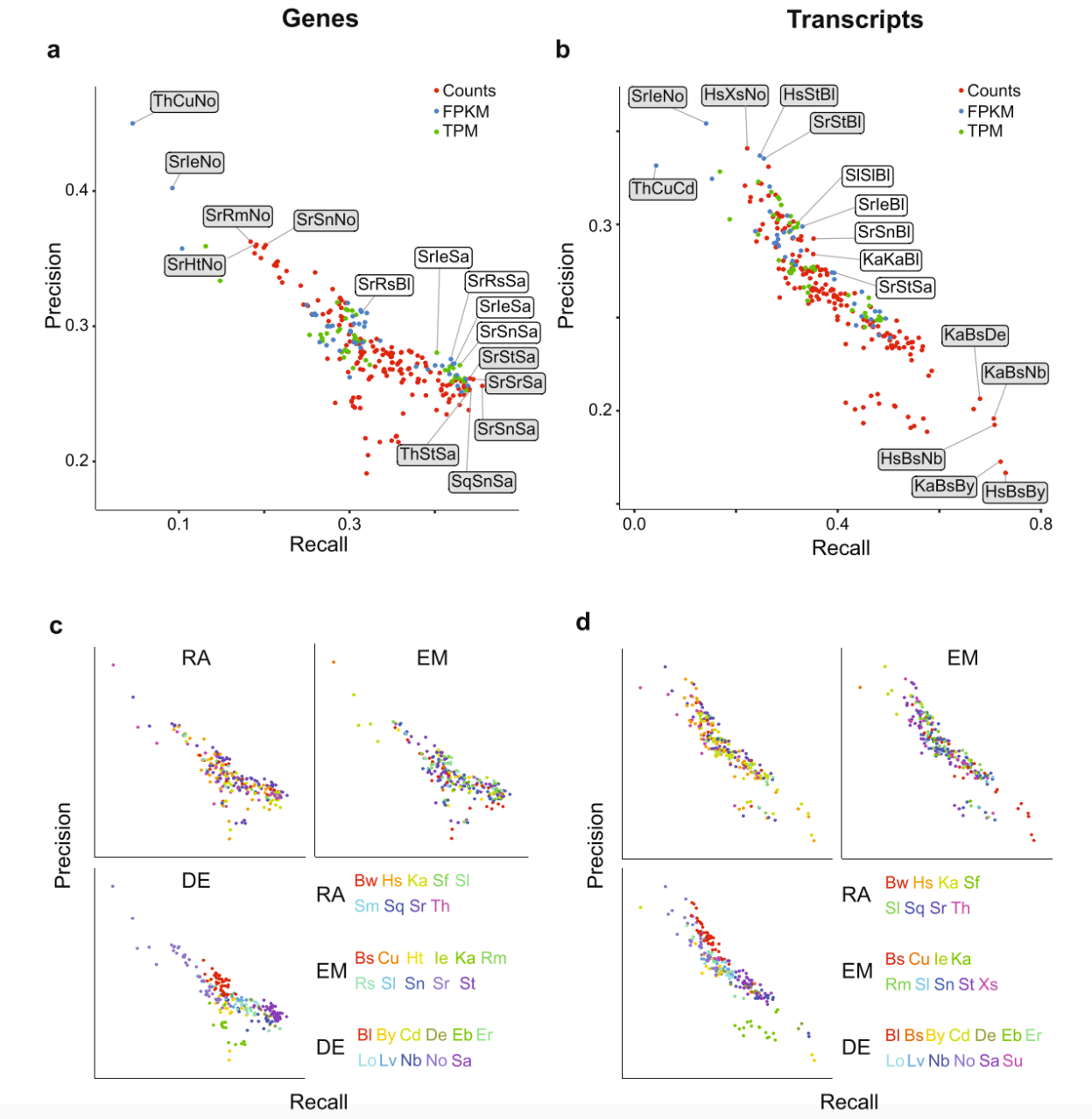

Figure 3.6 Comparison of performance metrics

(a, b) Precision and recall for each workflow, with top (shaded) and balanced (white) performers labeled. (c, d) Plots as above, with points colored by tool for each step.

3.10 Chapter Three Tables

Table 3.1 Analysis tools used in this study

Read Aligner	RA Code	Expression Modeler	EM Code	Differential Expression	DE Code
Bowtie2	Bw	BitSeq	Bs	Ballgown	Bl
HISAT2	Hs	cufflinks	Cu	BitSeq	Bs
Kallisto	Ka	htseq	Ht	baySeq	By
Salmon-FMD	Sf	IsoEM	Ie	cuffdiff	Cd
Sailfish	Sl	kallisto	Ka	DESeq2	De
SeqMap	Sm	RSEM	Rm	EBseq	Eb
Salmon-Quasi	Sq	rSeq	Rs	edgeR	Er
STAR	Sr	Sailfish	Sl	limma+voom	Lo
TopHat2	Th	Salmon	Sn	limma+vst	Lv
		STAR	Sr	NBPseq	Nb
		Stringtie	St	NOISeqBIO	No
		eXpress	Xs	SAMseq	Sa
				Sleuth	Su

Abbreviations specified in the table are used throughout the figures. Additional details are available in Additional file 1.

Chapter 4 : EMPIRICAL ASSESSMENT OF THE IMPACT OF SAMPLE NUMBER AND READ DEPTH ON RNA-SEQ ANALYSIS WORKFLOW PERFORMANCE

Chapter 4 is adapted with minimal modification from:

Baccarella A, Williams CR, Parrish JZ, Kim CC. (2018) Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinformatics19:423.

4.1 Abstract

4.1a Background

RNA-Sequencing analysis methods are rapidly evolving, and the tool choice for each step of one common workflow, differential expression analysis, which includes read alignment, expression modeling, and differentially expressed gene identification, has a dramatic impact on performance characteristics. Although a number of workflows are emerging as high performers that are robust to diverse input types, the relative performance characteristics of these workflows when either read depth or sample number is limited–a common occurrence in real-world practice–remain unexplored.

4.1b Results

Here, we evaluate the impact of varying read depth and sample number on the performance of differential gene expression identification workflows, as measured by precision, or the fraction of genes correctly identified as differentially expressed, and by recall, or the fraction of differentially expressed genes identified. We focus our analysis on 30 high-performing workflows, systematically varying the read depth and number of biological replicates of patient monocyte samples provided as input. We find that, in general for most workflows, read depth has little effect on workflow performance when held above two million reads per sample, with reduced workflow performance below this threshold. The greatest impact of decreased sample number is seen below seven samples per group, when more heterogeneity in workflow performance is observed. The choice of differential expression identification tool, in particular, has a large impact on the response to limited inputs.

4.1c Conclusions

Among the tested workflows, the recall/precision balance remains relatively stable at a range of read depths and sample numbers, although some workflows are more sensitive to input restriction. At ranges typically recommended for biological studies, performance is more greatly impacted by the number of biological replicates than by read depth. Caution should be used when selecting analysis workflows and interpreting results from low sample number experiments, as all workflows exhibit poorer performance at lower sample numbers near typically reported values, with variable impact on recall versus precision. These analyses highlight the performance characteristics of common differential gene expression workflows at varying read depths and sample numbers, and provide empirical guidance in experimental and analytical design.

4.2 Background

RNA sequencing (RNA-Seq), despite being the choice technique for transcriptome-wide identification of differentially expressed genes, is still rapidly evolving. A number of tools are available for each major processing step in data analysis: read alignment, expression modeling, and differential gene identification [1], but their performance in concert is only beginning to be understood [2–6]. Additionally, most existing evidence concerning their performance has been built on samples from simulated or highly controlled datasets, which lack the variability inherent in many experimental—particularly clinically-derived—datasets. We recently performed a broad comparison of RNA-Seq differential expression analysis workflows, applied to human clinical samples of highly purified monocyte subsets, using previously published microarray and BeadChip data as our reference gene sets. We found that a number of workflows performed poorly, but that the majority of workflows perform similarly well, with differences in their calibration with respect to having higher recall or precision [7].

Although the cost of RNA-Seq experiments has been steadily decreasing as technology improves, it continues to be an important consideration when designing RNA-Seq experiments. In particular, a large component of that cost is derived from the sample preparation and sequencing, where a tradeoff between read depth and sample number (*e.g.* the number of biological replicates per condition) must be taken into consideration [8]. Currently there is no consensus on adequate sequencing depth for

differential gene expression studies, and several studies on multiple organisms and sample types, varying in their heterogeneity, have suggested a wide range of read depths as optimal. At the low end, increasing read depths beyond 10 million reads was found to have minimal effect on the power to identify the effects of hormone treatment on a breast cancer cell line [8]; at the high end, increasing read depth up to 200 million reads, the highest depth available in the study, led to an increase, albeit small, in the detection of differentially expressed genes when comparing three male and three female human monocyte samples [9]. In simulations of three murine and four human data sets, including cell line, viral infection, disease, and population studies, a minimal power gain was seen above 20 million reads [10] or 30 million reads [11]. In studies employing one versus one sample comparisons, 200 million reads were suggested to identify the full range of differentially expressed transcripts between the MAQC dataset and a colorectal cancer line, and between human liver and kidney cells [12]; whereas 300 million reads was the depth identified as necessary to identify 80% of differentially expressed genes in human subcutaneous fat, pre and post induction of an inflammatory response [13].

Several studies have examined the interplay of sample number with read depth, with general consensus that increasing biological replicates increases power or gene recall more drastically than increasing read depth [8,10,14,15]. However, similar to read depth, recommendations for sample number vary. At the lower end of recommendations, three to four samples were determined to be sufficient for differential gene identification in a mouse neurosphere study, based on the relatively small incremental improvements in the AUC at these sample numbers [15], and three samples per group were necessary to detect the majority of differences tied to genotype, sex, and environment in a *Drosophila melanogaster* study, although additional replicates did increase power [16]. Ching *et al.* recommended a minimum of five samples per group, based on a variety of RNA-Seq datasets, but both they and others noted that much higher sample numbers are necessary to provide adequate power in samples with high gene dispersion, such as in a population comparison of Caucasian and Nigerian derived cells [10,15]. Similarly, in a *S. cerevisiae* study, a minimum of six replicates was recommended for differential expression studies, based on identification of true positive and false positive genes, although twelve replicates were necessary to identify the majority of differentially expressed genes [17]. It is likely that the wide range of organisms studied, including the genomic complexity, level of genetic heterogeneity within sample

groups, the magnitude of differences in conditions, and the target fraction of differences to be identified, impacts these disparate results across read depth and sample number. Beyond attempting to reach a generalization about input design, it is even harder to extrapolate these findings to clinically-derived samples. Several studies have proposed methods for sample size calculation [10,18–23]; however, their performance on real world data has shown widely variable results in sample size estimation, without clear indication that one outperforms the others [24]. As such, there remains a need for empirical studies of sample size effect, with a particular need for studies on experimental conditions more similar to study designs increasingly found in the literature.

Due to the paucity of real-world RNA-Seq samples for which reference datasets are available for comparison, studies of RNA-Seq tools have frequently been limited to using up-sampled RNA-Seq data as a metric of truth [8,10,13] or relied on highly controlled datasets generated *in silico* [10,11,15]. Additionally, many of the datasets frequently used for evaluation of RNA-Seq performance, such as the MAQC dataset [25] and the human kidney and liver dataset [26], do not allow for the study of effects of biological replicates, and exhibit extreme differences in gene expression that are not representative of typical study designs, where test groups are often much more closely related. Furthermore, many of the aforementioned studies only use a single workflow for differential gene identification, with none examining the interplay of tool choice at the three levels (read alignment, expression modeling, and differential expression) with read depth and sample number. Altogether, the limitations of these studies reflect the challenge of grappling with the multi-factorial parameters that can dramatically influence the performance of RNA-Seq analysis, and highlight the need for further assessments.

Here, we examined the effects of read depth and sample number on the performance of several differential expression analysis workflows, which have previously been identified as good performers when applied to input datasets with high read depth and sample number [7]. Performance was assessed using real-world, clinical samples of highly purified monocytes, with the use of four previously published microarray and BeadChip studies as a reference for biological truth [27–30]. The results of this study provide empirically-derived guidance to inform the design of RNA-Seq experiments, including the choice of RNA-Seq analysis workflow.

4.3 Methods

RNA-Seq Samples. The RNA-Seq datasets used in this study were previously published [7] and are available from the NCBI Sequence Read Archive (SRA) under accession number SRP082682.

Read subsampling, alignment, expression modeling, and differential expression identification. FASTQ files were randomly subsampled without replacement to create samples of depths $3x10^4$, $5x10^4$, $1x10^5$, $3x10^5$, $5x10^5$, $1x10^6$, $2x10^6$, $5x10^6$, $1x10^7$, and $2x10^7$ reads using the R package ShortRead [31], as allowed for by original read depth (Additional File 1). Each subset of reads was aligned to release GRCh38 of the human genome (Gencode Release 26) with HISAT2, Kallisto, Salmon, and STAR [32–35]. Gene expression was modeled with Kallisto, RSEM, Salmon, STAR, and Stringtie [33–37]. Gene counts obtained for genes on the pseudoautosomal region of the Y chromosome were excluded from further analysis, as they are identical in annotation and counts to these genes on chromosome X. For Kallisto and Salmon, transcript-level expression values were condensed to gene-level values using tximport [38]. Expression matrices for differential expression input were generated using custom scripts as well as the prepDE.py script provided at the Stringtie website. Ten iterations of differential expression analysis were run using different, randomly-chosen combinations of classical and nonclassical monocyte samples, with 3, 4, 5, 6, 7, 8, 9, 12, or 15 samples per group, using the same ten combinations for all workflows and read depths, as shown in Additional File 2. We kept sample combinations consistent for a given sample number when varying read depth to better isolate the effects of decreasing read depth. However, our sample combinations were restricted by the initial read depths of the samples (Additional File 1) and therefore some samples were excluded from individual analyses. Specifically, three samples were excluded from the two highest read depths (classical09, nonclassical01, and nonclassical10), and an additional nine samples were excluded from the highest read depth (classical01, classical04, classical10, classical15, classical16, nonclassical06, nonclassical07, nonclassical13, and nonclassical17). Because of these exclusions, we were unable to test 15 samples per group at the two highest read depths and 12 samples per group at the highest read depth. Differentially expressed genes were identified with Ballgown, DESeq2, edgeR exact test, limma coupled with voom transformation, NOISeqBIO, and SAMseq [39–44]. Of these, all but Ballgown and SAMseq used intrinsic filtering or

recommended extrinsic filtering of genes prior to testing. All differential expression tools were specified within the tool commands to run at a detection level of alpha of 0.05 or FDR of 0.05. In general, all software was run with default parameters; specific runtime parameters and software versions are listed in Additional File 3, and scripts for running all code are available at https://github.com/cckim47/kimlab/tree/master/rnaseq. Further information about implementation is available upon request. Note that tool and genome versions have been updated since our previous paper [7], so performance metrics may differ slightly.

Preparation of reference datasets. Reference datasets were prepared from four published studies of classical and nonclassical monocytes conducted on microarray or BeadChip platforms and retrieved from the NCBI Gene Expression Omnibus (GEO) with accession numbers GSE25913, GSE18565, GSE35457, GSE34515 [27–30], as previously described [7]. A third division of monocytes, intermediate monocytes, has recently been established [45], and these were isolated together with nonclassical monocytes in the two microarray experiments [28,30], but not in the Bead-Chip and RNA-Seq data sets [7,27,29]. Significant differentially expressed genes between classical and nonclassical monocytes were identified for each dataset with Significance Analysis of Microarrays (SAM) [46] with an FDR of 0.05, and limma [41], with a BH-adjusted p-value of 0.05. Performance of the RNA-Seq workflows against both the SAM and limma analyzed microarray data were previously compared to one another and found to exhibit good reproducibility regardless of the statistical method used to generate the microarray data [7]; as such, we chose to use the genes at the intersection of the two methods for our final reference gene sets here.

Quantification of recall and precision. Both performance metrics were calculated as previously described [7]. Because absolute recall and precision values are influenced by the repertoire of analytes that can be measured by a given platform, we first filtered each reference and RNA-Seq gene set to include only features measurable both by RNA-Seq (*i.e.*, present in the GRCh38 genome release, Gencode version 26) and by the microarray (*i.e.*, a probe targeting the feature was present on the microarray platform) within a given comparison. All gene set counts are reported based on these filtered numbers, as are all calculations of recall and precision. Recall was calculated as the number of significant genes in the

intersection of the test RNA-Seq dataset with the reference dataset, divided by the number of genes identified as significant in the reference dataset. Precision was calculated as the number of significant genes in the intersection of the test RNA-Seq dataset with the reference dataset, divided by the number of genes identified as significant in the test RNA-Seq dataset.

Literature survey. To survey current sample number practices in the RNA-Seq literature, the following PubMed search was queried in January 2018: ((rna seq OR rna-seq OR RnaSeq)) AND (differential OR differentially) NOT (miRNA OR non-coding OR "single cell" OR lncRNA OR "circular RNA"). There was no date-based selection, but the earliest studies in the randomly chosen datasets were from 2010. Results were either unfiltered or filtered on species “human”. Only studies which performed differential expression analysis were included. Studies utilizing previously published datasets, including large-scale sequencing efforts (such as TCGA) were excluded, to ensure a representative sampling of the most common experimental designs. For the human-specific survey, only studies utilizing primary patient samples or cell lines derived from individual patients were included. Results of either filtered or unfiltered searches were randomized, and then reviewed sequentially until 100 papers meeting inclusion criteria were reviewed. Average sample number was determined by adding the number of samples included in each pairwise comparison, divided by two times the total the number of comparisons in a given study (Additional File 4). Sample number was corroborated by two authors for 10% of papers with 90% concordance between two reviewers. In studies in which there was discordance in the average sample number calculated, both authors independently reviewed the study and were in agreement following this review.

4.4 Results

4.4a Generation of subsampled real-world RNA-Seq dataset for benchmarking

We sought to empirically assess the impact of read depth and sample number on RNA-Seq workflow performance, using patient-derived clinical samples, which integrate many sources of variability that are not well represented in typical benchmarking datasets. Our RNA-Seq dataset has been previously described [7]. In brief, RNA from nonclassical and classical monocytes was isolated from

cryopreserved PBMCs collected as part of a study of Ugandan children. RNA-Seq libraries were sequenced as 51-base single-end reads on an Illumina HiSeq 2500. Total reads per sample were variable, ranging from 6 to 37 million, but with no significant difference of read number or quality between the 16 classical and 16 nonclassical samples [7]. Raw fastq files were randomly subsampled to create fastq files of $3x10^4$, $5x10^4$, $1x10^5$, $3x10^5$, $5x10^5$, $1x10^6$, $2x10^6$, $5x10^6$, $1x10^7$, and $2x10^7$ reads for each sample, as allowed for by the original read depth (Additional File 1).

4.4b Overview of Empirical Testing

As previously described, we used four studies which explored expression differences between classical and nonclassical monocytes, using microarray and BeadChip analysis [27–30], to generate a reference of biological truth for comparison. By utilizing data from four independent studies we were able to minimize the effect that any individual preparation had on the results, while still analyzing clinical samples with inter-sample variability and effect size characteristics commonly found in real-world RNA-Seq studies, but not in traditional benchmarking datasets. Despite differences in collection and processing methods, as well as variability in the genetic backgrounds between studies, in our previous analysis we found that the performance of various RNA-Seq workflows was remarkably consistent when using any of the four reference datasets as truth [7]. As such, we have chosen to report only on performance averaged across the four datasets for the current analysis. Additionally, we previously found strong concordance between results when using either SAM or limma to detect differentially expressed genes from the microarray and BeadChip datasets [7], so have used the intersection of the two analysis methods to generate our "ground-truth" gene lists.

With these four datasets as our references for performance comparisons, we focused our evaluation of RNA-Seq analysis workflows on those which we had previously identified as “high performers” -- high recall, high precision, or among the top in combined performance [7]. From within this subset, we additionally selected for commonly used workflows, ease of implementation, and run speed. Finally, to constrain our exploration space, we limited our analysis to workflows that would ultimately lead to differential expression testing done on read counts (as opposed to FPKM or TPM), and on gene level data (as opposed to transcript level). Based on our previous analysis [7] that found that the differential

expression analysis tool had the largest effect on performance, we predicted that it would be more informative to include more differential expression detectors than read aligners or expression modelers. In total, we evaluated four read aligners, five expression modelers, and six differential expression detectors, as shown in Table 4.1, in thirty total combinations. We applied these workflows to ten iterations of randomly selected samples at each of the following number of samples per group: 3, 4, 5, 6, 7, 8, 9, 12, and 15. Because only 17 samples per condition were available, the higher sample number iterations had more overlapping samples than the lower sample number iterations. The same 10 sample iterations were used at all read depths and across all workflows for consistency among comparisons. If there were not enough samples available for 10 distinct sample combinations at a given read depth, the sample number/read depth combination was not run. To benchmark performance, we calculated the precision (intersecting significant genes divided by total number of significant genes identified by RNA-Seq) and recall (intersecting significant genes divided by the total number of significant reference genes) of each iteration.

4.4c Differential influence of workflow stages

For each workflow consisting of all three steps (read alignment, expression modeling and differentially expressed gene identification), we evaluated the ability to detect genes differentially expressed between classical and nonclassical monocytes, at each aforementioned read depth and sample number, over 10 iterations of sample combinations at each sample number. We first wanted to examine the relative impact of each step on precision and recall, over the search space. Strikingly, it was visually clear that the choice of differential expression tool was much more impactful than the choice of the read aligner/expression modeler pair, with performance tending to cluster by differential expression tool over iterations at each read depth and sample number (Figure 4.1 and Additional File 5; figures show the same data with shape and color labels inverted). For example, Ballgown showed a large range of precision values across all read alignment and expression modeling pairs whereas NOISeqBio consistently exhibited a large range in recall values. This is similar to our previous finding that the choice of differential expression tool was the most impactful workflow step, in the absence of read or sample subsampling [7]. Given the similarity in performance across different upstream steps when holding

differential expression tool constant, we have chosen to subsequently highlight the differential expression tool rather than complete workflow in figures, for ease of comparison between differential expression tools. To facilitate a more in-depth exploration of the data, we also provide an interactive figure that enables visualization of performance metrics for individual workflows (Additional File 6).

We note that we have compared significant gene lists to each of four microarray datasets individually and then calculated an average performance across the datasets. Since any two of the truth datasets exhibited at least 500 unique differentially detected significant genes in a direct comparison [7], it is not surprising that absolute precision was not high when calculated with each truth dataset, and supports the hypothesis that there would be variability across independently collected and analyzed patient monocyte samples.. It is likely that this and other factors play a combinatorial role in explaining the low absolute precision values and demonstrate the difficulty of defining a ground truth for genome-wide studies. As such, we advise focusing on the relative comparisons of various workflows' precision/recall trade-off which provides useful guidance when weighing options for RNA-Seq analysis.

4.4d Effects of read depth on performance

Within each workflow, performance varied dramatically as read depth and sample number per group varied. To isolate the effect of read depth on the precision and recall of the workflows, we focused on iterations run on the highest numbers of samples per group to minimize the impact of sample number effect on interpretation. At the higher read depths, a performance trade-off between precision and recall was present when comparing workflows, following an inverse linear relationship (Figure 4.2), as we previously reported [7]. To aid in visual comparison of performance at each read depth to the original, highest read depth performance, we have depicted the original regression line at each subsequent sub-sampled read depth. This inverse linear relationship degrades as read depth decreases, primarily due to a loss in recall. Initial degradation in linearity becomes apparent at $2x10^6$ reads, with a drop in correlation, and the majority of workflows deviate from the high-read regression line by $1x10^5$ reads (Figure 4.2 and Additional File 7). This was seen consistently at sample numbers of 15, 12 and 9, with only slight variation in the pattern of degradation at the different sample numbers. Throughout the range of read depths, differential expression tools largely maintain their precision and recall positions relative to other tools

(Additional File 7 and Additional File 8), although Ballgown's recall with nine samples more quickly degraded than the other tools as read depth decreased. Limma-voom and edgeR coupled with HISAT2-Stringtie also lost recall more rapidly than when coupled with the other read aligner / expression modeler combinations, at all three sample numbers (Figure 4.2 and Additional File 6). At these higher sample numbers, NOISeqBIO was comparatively resilient to effects of decreasing read depth, with maintenance of its balance between precision and recall across the tested depths (Additional File 6). Of note, SAMseq was unable to consistently handle the lower read depths, with the majority of iterations failing at $1x10^5$ reads and all iterations failing at $3x10^4$ read depth (Additional File 9). SAMseq's performance improved as read depth increased, and was able to run with no failures at the two highest read depths (Additional File 9). While the inner workings of any individual tool are beyond the scope of this study, it appears that the failures were due to errors following SAMseq's read depth estimation, when the function samr.estimate.depth() returns an estimated read depth of 0 for at least one sample in each failed comparison. Given that initial estimates for RNA-Seq read depth requirements were estimated to be one to two orders of magnitude higher than the failure depths [12,13], it is likely SAMseq was not developed to handle these lower read depths. Of note, increasing sample size does decrease the number of failures from SAMseq for the intermediate-high read depths (Additional File 9), suggesting an interaction between these parameters.

4.4e Effects of sample number on performance

Decreasing read depths consistently led to decreased performance—particularly decreased recall—at high sample numbers; however, it was clear that the slope of the recall-precision relationship shifted as sample number changed (Figure 4.2). To more closely examine the effects of sample number, we limited examination to the highest read depths. Surprisingly, reduction in sample number was impactful from a relatively high number of samples. Changes in the slope of the recall-precision relationship were apparent from eight samples per group, with a large decrease in the recall-precision linearity relationship at six samples per group (Figure 4.3). Similar to our findings with read depth, this change was most reflected by loss of recall, disproportionate to the loss of precision (Figure 4.3 and Additional File 7). Notably, Ballgown performed particularly poorly at the lowest sample numbers, with

many iterations failing to call any significant genes, leading to precision and recall values of zero (Figure 4.3). This poor performance was not related to the upstream choice of read aligner and expression modeler, but rather the same sample groupings tended to return poor results for all upstream workflow combinations. As noted by the Ballgown authors, although Ballgown shares a similar underlying linear model to limma for identification of differentially expressed genes, the initial empirical Bayes modeling employed by limma prior to differential testing allows for shrinkage of variance estimates, which has a larger effect for smaller sample sizes where less biological replicate information is available [44]; thus, limma has superior performance at low sample numbers, as we see here. NOISeqBIO also demonstrated unusual behavior at the lowest sample numbers – at three and four samples per group, performance skewed heavily towards recall, with very low precision, the opposite of the performance seen at higher sample numbers. This behavior was independent of read depth (Figure 4.3 and Additional File 7). NOISeqBIO combines a non-parametric model with an empirical Bayes approach to shrink variance estimates, regardless of sample number. However, when NOISeqBIO is used with fewer than five samples per group, there is a change in the methodology for the creation of the null distribution of its non-parametric model. At these lower sample numbers, k-means clustering is employed to identify genes with similar expression patterns and information is shared between these genes when creating the null distribution, whereas this is not done at higher sample numbers [42]. This difference in methodology for managing lower sample numbers might explain the abrupt shift to high recall with reduced precision.

Given the somewhat surprising result that sample numbers below six had severely reduced performance, we next sought to assess how widely used low sample numbers were in recent RNA-Seq studies. From 100 randomly chosen studies, over 90% used six or fewer samples per group (Figure 4.4A). When this survey was repeated and restricted to studies of human samples, the average sample numbers were slightly higher, with about half of the studies falling at or below six samples per group (Figure 4.4B). This suggests that while some authors of human studies recognize the increased variability inherit to clinical samples and increase sample size accordingly, the performance characteristics of many human studies would be improved with increased sample numbers. Given these results, caution should be exercised in interpreting many recent RNA-Seq studies that may conform to common experimental design approaches, but that may be underpowered for RNA-Seq analysis. Additionally, this highlights the

necessity of benchmarking RNA-Seq tool performance using datasets most similar to those that the methods will be applied to, to better define best practices for study design and analysis.

4.4f Correlations with significant gene numbers

In our initial study examining the performance of workflows, we observed that the number of genes called significant by a workflow heavily influenced the recall and precision, with a strong correlation between recall and the number of genes identified as significant, and an inverse relationship between precision and the number of significant genes [7] . As such, we hypothesized that changes in the number of genes identified as significant would be correlated with the degradation of performance at lower sample numbers and read depths. As predicted, we observe a strong relationship between recall and number of genes called significant, with the number of genes called significant tending to increase as sample number increased, with a commensurate increase in recall (Figure 4.5A and Additional File 10). Surprisingly, and in contrast to our previous observations across workflows [7], the converse was not true for precision (Figure 4.5B). While the trend that higher numbers of genes called significant tended to have lower precision remained true, this effect was much less pronounced. Interestingly, the precision across workflows tended to decrease at the highest sample numbers. While this could represent an increase in falsely called genes as the total number of significant genes increases, it is also possible that at these higher sample numbers RNA-Seq overtakes microarray's and BeadChip's abilities to detect differentially expressed genes. Notably, workflows employing NOISeqBIO at three and four samples called the highest number of significant genes of any workflows, which likely accounts for the relatively high recall with poor precision displayed by these workflows at low sample numbers. This suggests that results from NOISeqBIO must be interpreted with caution at low sample numbers, due to the higher likelihood of type I error.

4.5 Conclusions

Of the workflows examined, all performed well at higher read depths and sample numbers, and the choice of workflows at these parameters should be largely influenced by the tolerance of a specific

application for type I versus type II error, as we concluded previously [7]. However, caution should be used at lower read depths and sample numbers, as performance is variable and highly dependent on the choice of differential expression tool, with much smaller impact from read aligner and expression modeler. These results also give insight into the read depth and sample number required for robust results when designing RNA-Seq experiments involving clinical samples, which exhibit more genetic and pre-analytical heterogeneity than typical *in vitro* study designs. Performance was relatively resistant to changes in read depth, with very minimal impact down to two million reads, which is considerably lower than previously published suggested read depths and may reflect analysis of different organisms and/or sample types [10–13]. Conversely, tool performance—particularly recall and the commensurate number of genes called significant—rapidly declined as sample number per group decreased, with changes apparent even by eight samples. At six or fewer samples per group, tool choice became increasingly impactful with SAMseq and Ballgown falling below the linear relationship, and thus being "worse" performers in this context. These findings corroborate past suggestions that increasing biological replicates will generally have a greater impact than increasing read depth [8,10,14,15], although this also depends on the "starting points" for read depth and sample number per group. If sample number is constrained, caution must be exercised in choosing a differential expression tool, as performance is more variable. Specifically, there is increased risk of Type II error, most disproportionately when using Ballgown with the lowest sample numbers, and Type I error in the case of NOISeqBIO used with fewer than five samples. These findings represent a departure from current practices used in many studies, which tend to follow more traditional experimental designs employing fewer replicates.

4.6 Declarations

List of Abbreviations Used

RNA-Seq: RNA sequencing

PBMCs: peripheral blood mononuclear cells

FDR: false discovery rate

FPKM: fragments per kilobase of transcript per million mapped reads

TPM: transcripts per million

Acknowledgements

Not applicable.

Funding

This work was supported by a grant from the National Institutes of Health, University of California, San Francisco-Gladstone Institute of Virology & Immunology Center for AIDS Research, P30 AI027763, NIAID U19 AI089674, NIAID R21 AI114916, NEI U10 EY008057, and NIDDK P30 DK063720 to CCK; a National Institutes of Health grant NINDS R01 NS076614 and a UW Research Innovation award to JZP; an NSF Graduate Research Fellowship (DGE1256032) and a National Institute of Health grant NINDS F31- NS106775-01 to CRW; and an ACCMA Community Health Foundation Summer Scholarship and a Schoeneman Scholarship to AB. None of the funding bodies played a role in the design of the study; collection, analysis, or interpretation of the data; or writing of the manuscript.

Availability of data and materials

The human monocyte RNA-Seq data set supporting the conclusions of this article is available in the NCBI Sequence Read Archive (SRA) under accession number SRP082682, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA339968. The monocyte microarray data sets supporting the conclusions of this article were obtained from the NCBI Gene Expression Omnibus (GEO) under accession numbers GSE25913, GSE18565, GSE35457, GSE34515, found at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25913, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18565, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35457, and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34515.

Authors' contributions

AB performed the sample number and read depth iteration analysis. AB, CRW, and CCK performed the published workflow analysis. AB, CRW, JZP, and CCK wrote the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

4.7 Supplemental Information

Additional files referenced in the text of this chapter are available with the published version of this manuscript, found at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2445-2. The content of these files is as follows:

Additional File 1: Original read depth of individual samples.

Format: XLSX Size: 13 kB

This file can be viewed with: Microsoft Excel

Additional File 2: Sample combinations for each iteration at varying sample numbers. The same sample combinations were run at all read depths and for all workflows.

Format: XLSX Size: 22 kB

This file can be viewed with: Microsoft Excel

Additional File 3: Table of software tools, with versions and runtime parameters.

Format: XLSX Size: 16 kB

This file can be viewed with: Microsoft Excel

Additional File 4: Literature survey citations and average sample number. 200 studies containing RNA-Seq differential expression analysis, either from all species or limited to primary human samples. Average sample number from these studies is also displayed in Figure 4.

Format: XLSX Size: 53 kB

This file can be viewed with: Microsoft Excel

Additional File 5: Analysis workflow steps impact on performance. Precision and recall for each iteration, separated by read aligner and expression estimator (rows) and differential gene tool (columns). Colors represent read depths and shapes represent sample number. These are the same data presented in Fig 1 with color and shape labels switched.

Format: PDF Size: 2301 kB

This file can be viewed with: Adobe Acrobat Reader

Additional File 6: Interactive figure for comparison of performance metrics. (A) Absolute precision and recall for each workflow. (B) Relative ranks of precision and recall for each workflow. Grey dots show performance for all workflows for the selected read depth(s) and sample number(s); red dots highlight the selected workflow(s).

Format: XLSX Size: 537 kB

This file can be viewed with: Microsoft Excel

Additional File 7: Impact on performance by read depth and sample number.

Precision and recall, averaged over the 10 iterations at a given sample number and read depth, split by sample number (columns) and read depth (rows). Values for each workflow (read aligner, expression modeler, and differential expression tool) are averaged and displayed separately. Points represent mean;

bars represent standard deviation; colors represent differential expression tool. Red line represents Lm fit for plotted data. Text is the corresponding R^2 value.

Format: PDF Size: 9557 kB

This file can be viewed with: Adobe Acrobat Reader

Additional File 8: Impact on rank performance by read depth and sample number.

Rank precision and rank recall, averaged over the 10 iterations at a given sample number and read depth, split by sample number (columns) and read depth (rows). Values for each workflow (read aligner, expression modeler, and differential expression tool) are averaged and displayed separately. Points represent mean; bars represent standard deviation; colors represent differential expression tool. Red line represents Lm fit for plotted data. Text is the corresponding R^2 value.

Format: PDF Size: 9676 kB

This file can be viewed with: Adobe Acrobat Reader

Additional File 9: Number of SAMseq failed iterations

Iteration was counted as a failure if SAMseq was not successfully run due to an error message. Bars represent count of failures, colored by read depth.

Format: PDF Size: 143 kB

This file can be viewed with: Adobe Acrobat Reader

Additional File 10: Number of significant genes by number of biological replicates.

Bar represents average number of significant genes for a given read depth, sample number, and differential expression tool. Average was calculated by averaging each of the ten sample combination iterations at a given sample number and read depth, for all five read aligner/expression modeler combinations upstream of a given differential expression tool. Standard deviation is shown. Colored by read depth.

Format: PDF Size: 785 kB

This file can be viewed with: Adobe Acrobat Reader

4.8 References

1. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. Genome Biol. BioMed Central; 2010;11:220.

2. Fonseca NA, Marioni J, Brazma A, Salzberg S, Zaleski C. RNA-Seq Gene Profiling - A Systematic Empirical Comparison. Provart NJ, editor. PLoS One. Public Library of Science; 2014;9:e107026.

3. Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Res. Oxford University Press; 2012;40:10084–97.

4. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16:59–70.

5. Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. BioMed Central; 2016;17:74.

6. Merino GA, Conesa A, Fernández EA. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. Brief Bioinform. 2017;

7. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics. 2017;18:38.

8. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. W. H. Freeman, New York; 2014;30:301–4.

9. Mirsafian H, Ripen AM, Leong W-M, Manaharan T, Mohamad S Bin, Merican AF. Transcriptome landscape of human primary monocytes at different sequencing depth. Genomics. 2017;

10. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA. Cold Spring Harbor Laboratory Press; 2014;20:1684–96.

11. Williams AG, Thomas S, Wyman SK, Holloway AK, Williams AG, Thomas S, et al. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Curr Protoc Hum Genet. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2014. p. 11.13.1-11.13.20.

12. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21:2213–23.

13. Liu Y, Ferguson JF, Xue C, Silverman IM, Gregory B, Reilly MP, et al. Evaluating the Impact of Sequencing Depth on Transcriptome Profiling in Human Adipose. Liu Z, editor. PLoS One. Public Library of Science; 2013;8:e66883.

14. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14:R95.

15. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One. Public Library of Science; 2014;9:e103207.

16. Lin Y, Golovnina K, Chen Z-X, Lee HN, Negron YLS, Sultana H, et al. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. BMC Genomics. BioMed Central; 2016;17:28.

17. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. Cold Spring Harbor Laboratory Press; 2016;22:839–51.

18. Wu H, Wang C, Wu Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. Bioinformatics. 2015;31:233–41.

19. Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics. Oxford University Press; 2013;29:656–7.

20. Bi R, Liu P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. BMC Bioinformatics. BioMed Central; 2016;17:146.

21. Zhao S, Li C, Guo Y, Sheng Q, Shyr Y. RnaSeqSampleSize: RnaSeqSampleSize. R package; 2017.

22. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. Calculating sample size estimates for RNA sequencing data. J Comput Biol. 2013;20:970–8.

23. van Iterson M, van de Wiel MA, Boer JM, de Menezes RX. General power and sample size calculations for high-dimensional genomic data. Stat Appl Genet Mol Biol. De Gruyter; 2013;12:449–67.

24. Poplawski A, Binder H. Feasibility of sample size calculation for RNA-seq studies. Brief Bioinform. 2017;bbw144.

25. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. BioMed Central; 2010;11:94.

26. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18:1509–17.

27. Haniffa M, Shin A, Bigley V, McGovern N, Teo P, See P, et al. Human tissues contain CD141hi cross-presenting dendritic cells with functional homology to mouse CD103+ nonlymphoid dendritic cells. Immunity. 2012;37:60–73.

28. Frankenberger M, Hofer TPJ, Marei A, Dayyani F, Schewe S, Strasser C, et al. Transcript profiling of CD16-positive monocytes reveals a unique molecular fingerprint. Eur J Immunol. 2012;42:957–74.

29. Wong KL, Tai JJ-Y, Wong W-C, Han H, Sem X, Yeap W-H, et al. Gene expression profiling reveals the defining features of the classical, intermediate, and nonclassical human monocyte subsets. Blood. 2011;118:e16–31.

30. Ingersoll M a., Spanbroek R, Lottaz C, Gautier EL, Frankenberger M, Hoffmann R, et al. Comparison of gene expression profiles between human and mouse monocyte subsets. Blood. 2010;115:e10-19.

31. Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics. Oxford University Press; 2009;25:2607–8.

32. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. Nature Publishing Group; 2015;12:357–60.

33. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

34. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. Nature Publishing Group; 2017;14:417–9.

35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

36. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 2011;21:940–51.

37. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. Nature Publishing Group; 2015;33:290–5.

38. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4:1521.

39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. BioMed Central Ltd; 2014;15:550.

40. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.

41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47–e47.

42. Tarazona S, Furió-Tarí P, Turrà D, Pietro A Di, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. Oxford University Press; 2015;43:e140.

43. Li J, Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22:519–36.

44. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. NIH Public Access; 2016;11:1650–67.

45. Ziegler-Heitbrock L, Ancuta P, Crowe S, Dalod M, Grau V, Hart DN, et al. Nomenclature of monocytes and dendritic cells in blood. Blood. 2010;116:e74–80.

46. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98:5116–21.

47. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol. 2015;33:243–6.

48. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. BioMed Central; 2014;15:R29.

4.9 Chapter Four Figures

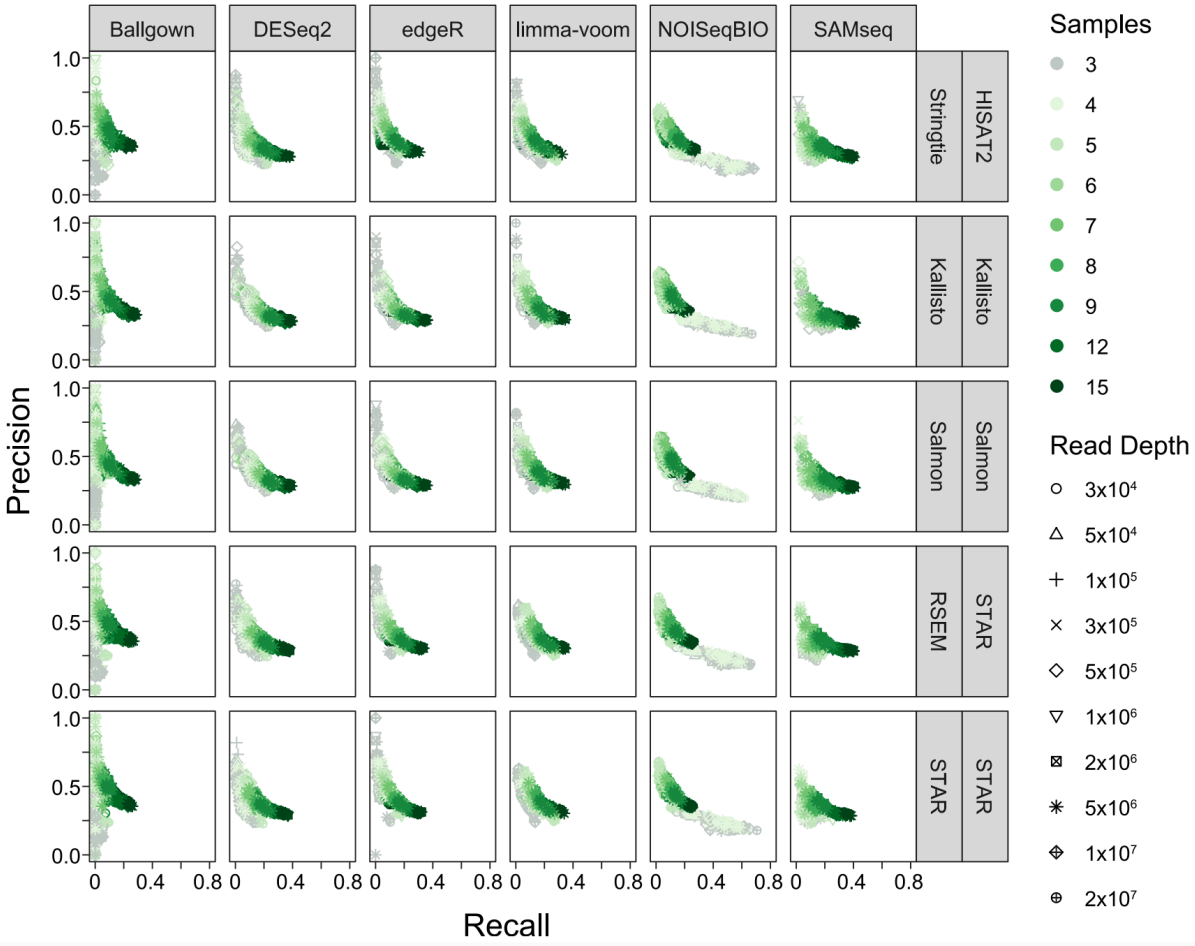

Figure 4.1 Analysis workflow steps' impact on performance

Precision and recall for each iteration, separated by read aligner and expression modeler (rows) and differential gene tool (columns). Colors represent sample number and shapes represent read depths.

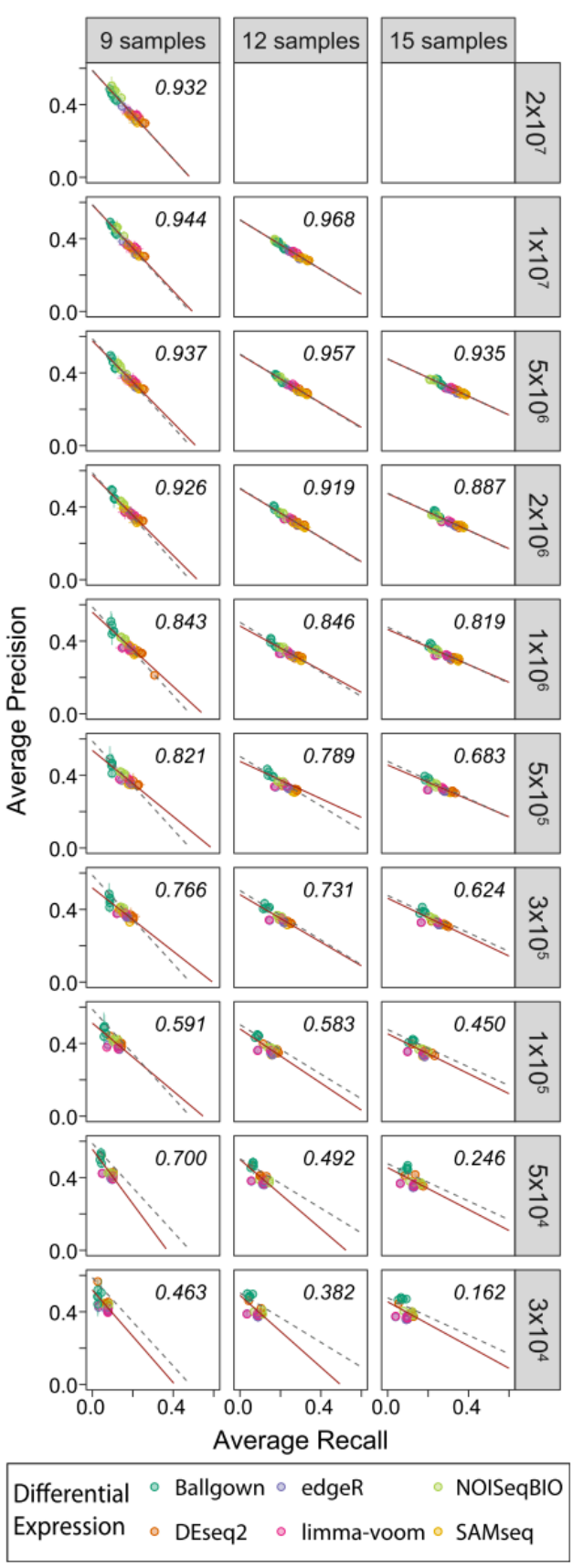
9 samples
12 samples
15 samples
2x10^7
1x10^7
5x10^6
2x10^6
1x10^6
5x10^5
3x10^5
1x10^5
5x10^4
3x10^4
0.932
0.944
0.968
0.937
0.957
0.935
0.926
0.919
0.887
0.843
0.846
0.819
0.821
0.789
0.683
0.766
0.731
0.624
0.591
0.583
0.450
0.700
0.492
0.246
0.463
0.382
0.162
Average Precision
Average Recall
0.0
0.4
Differential Expression
Ballgown
edgeR
NOISeqBIO
DEseq2
limma-voom
SAMseq

Figure 4.2 Read depth's impact on performance

Precision and recall, averaged over the 10 iterations at a given sample number and read depth, split by sample number (columns) and read depth (rows). Values for each workflow (read aligner, expression modeler, and differential expression tool) are averaged and displayed separately. Points represent mean; bars represent standard deviation; colors represent differential expression tool. Red solid line represents linear regression line for plotted data. R^2 value corresponds to plotted data. Gray dashed line represents linear regression fit of the first row of data for each column, superimposed over subsequent rows for comparison.

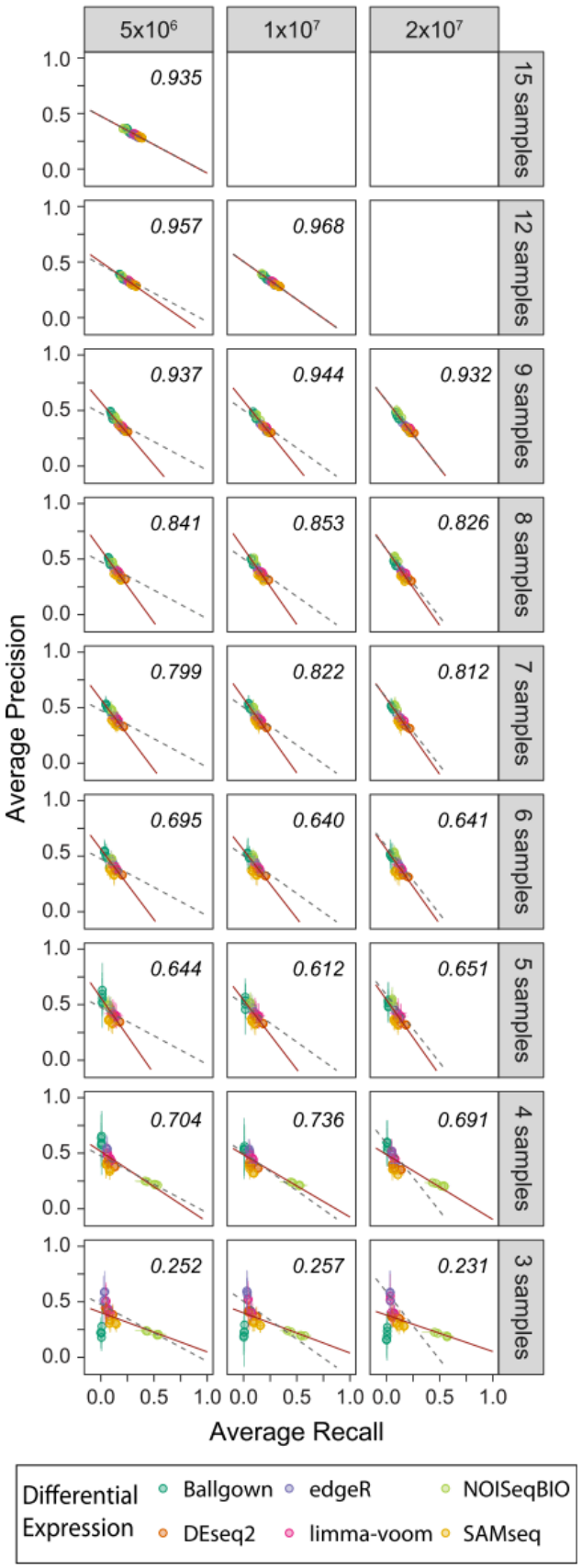
5x10^6
1x10^7
2x10^7
15 samples
12 samples
9 samples
8 samples
7 samples
6 samples
5 samples
4 samples
3 samples
0.935
0.957
0.968
0.937
0.944
0.932
0.841
0.853
0.826
0.799
0.822
0.812
0.695
0.640
0.641
0.644
0.612
0.651
0.704
0.736
0.691
0.252
0.257
0.231
Average Precision
Average Recall
0.0
0.5
1.0
Differential Expression
Ballgown
edgeR
NOISeqBIO
DEseq2
limma-voom
SAMseq

Figure 4.3 Sample number's impact on performance

Precision and recall, averaged over the 10 iterations at a given sample number and read depth, split by read depth (columns) and sample number (rows). Values for each workflow (read aligner, expression modeler, and differential expression tool) are averaged and displayed separately. Points represent mean; bars represent standard deviation; colors represent differential expression tool. Red solid line represents linear regression line for plotted data. R^2 value corresponds to plotted data. Gray dashed line represents linear regression fit of the first row of data for each column, superimposed over subsequent rows for comparison.

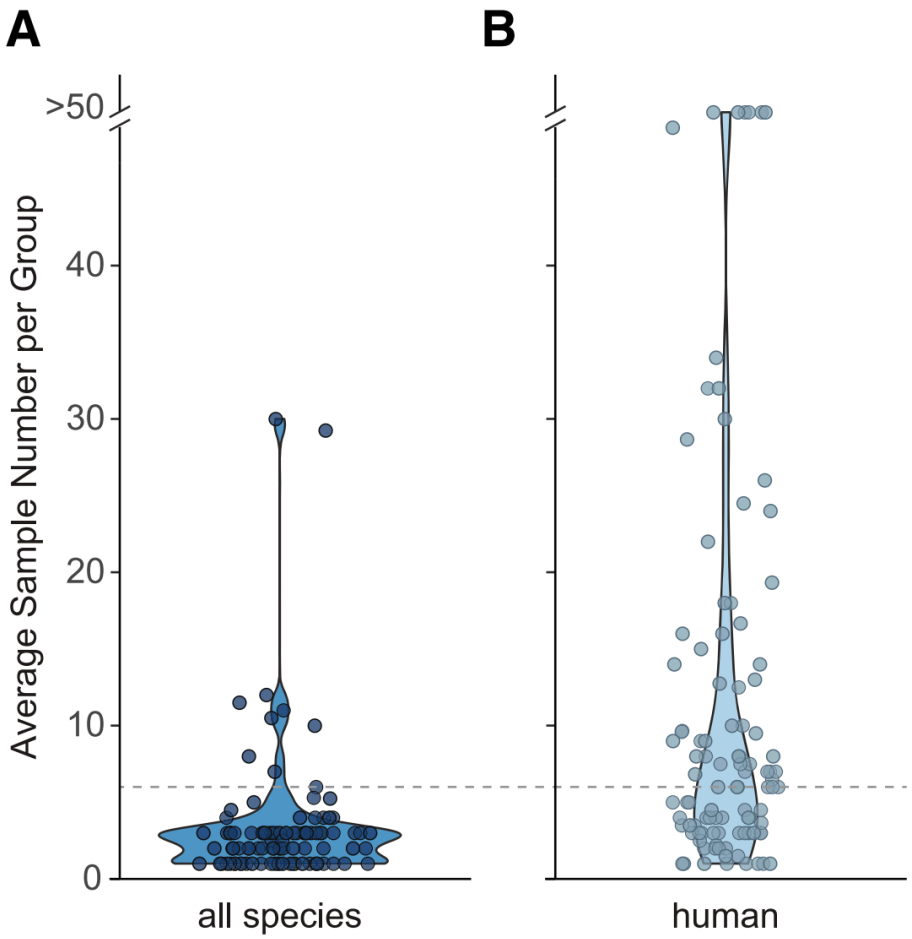

Figure 4.4 Literature survey of RNA-Seq experiment sample numbers

Violin plots of sample numbers used in 200 studies containing RNA-Seq differential gene expression analysis, either from all species (A) or limited to primary human samples (B). Individual dots represent average sample number used in each study. Grey dashed line represents six samples.

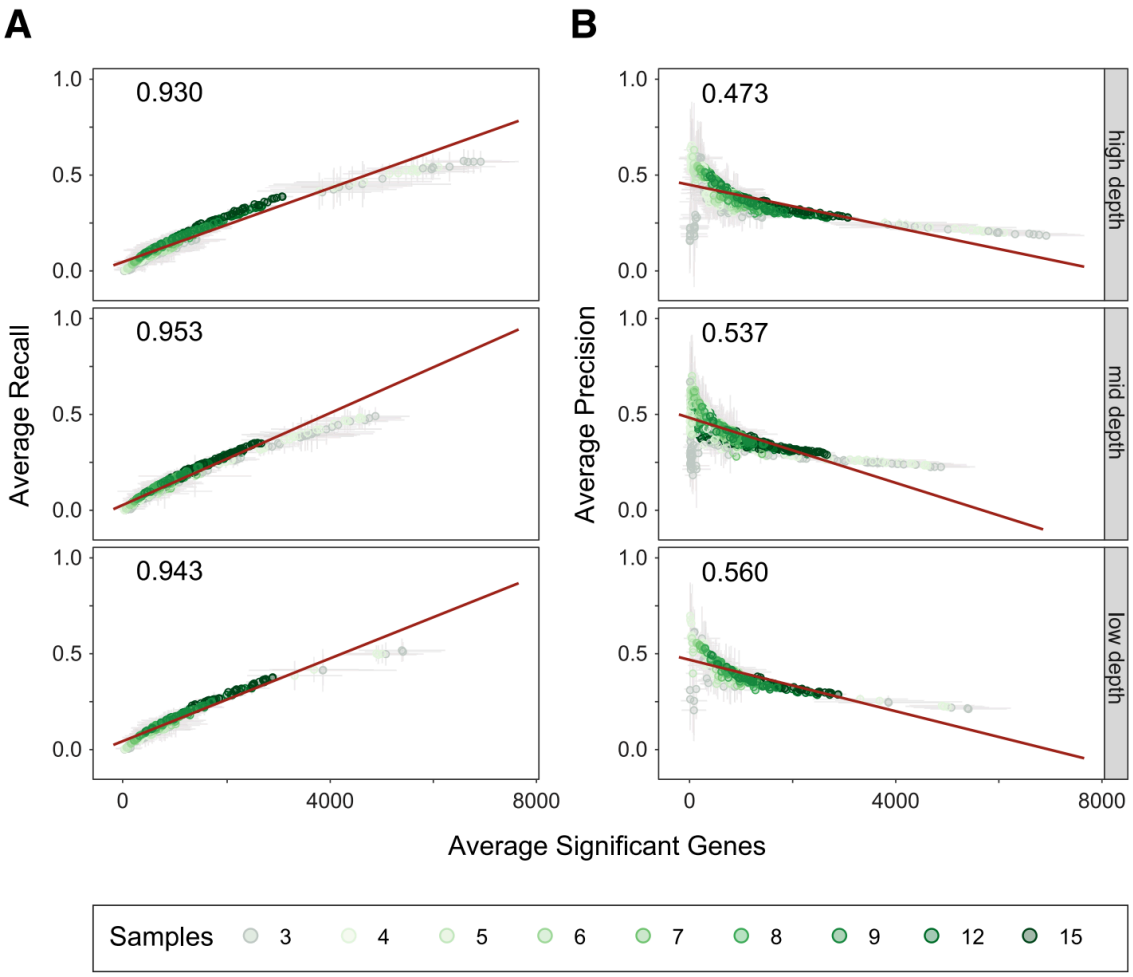

Figure 4.5 Significant gene number's impact on performance

Average recall (A) or average precision (B) versus the average number of genes identified as significant. Panels are split by read depths, with $2x10^7$, $1x10^7$, $5x10^6$, and $2x10^6$ reads plotted as high read depths, $1x10^6$, $5x10^5$ and $3x10^5$ plotted as medium read depths, and $1x10^5$, $5x10^5$, and $3x10^4$ plotted as low read depths. Dots represent values for individual workflows (read aligner, expression modeler, and differential expression tool) at a given sample number and read depth, averaged over the ten sample combination iterations run at each given sample number and read depth. Bars represent standard deviation. Colors represent sample number. Red line represents linear regression for plotted data. R^2 value corresponds to plotted data.

4.10 Chapter Four Tables

Table 4.1 Analysis tools used in this study

Read Aligner / Expression Modeler	Differential Expression Tool
HISAT2 / Stringtie	Ballgown
Kallisto / Kallisto	DESeq2
Salmon / Salmon	edgeR
STAR / RSEM	limma-voom
STAR / STAR	NOISeqBio
	SAMseq

Additional details are available in Additional File 3.

Chapter 5 : SELECT APPLICATIONS OF RNA-SEQ ANALYSIS TECHNIQUES

In parallel to establishing guidelines for best practices in RNA-Seq analysis, I collaborated on multiple projects that utilized RNA-Seq to address specific biological questions. In this chapter, I present selections from three such projects for which I collected and / or analyzed RNA-Seq data.

5.1 TrpA1 expression in nociceptors underlies chemotherapeutic-induced pain hypersensitivity.

The following selection is adapted with minimal modification from:

Boiko N, Medrano G, Montano E, Jiang N, Williams CR, Madungwe NB, Bopassa JC, Kim CC, Parrish JZ, Hargreaves KM, Stockand JD, Eaton BA. (2017) TrpA1 activation in peripheral sensory neurons underlies the ionic basis of pain hypersensitivity in response to vinca alkaloids. PLoS ONE 12(10): e0186888.

5.1a Abstract

Chemotherapy induced peripheral neuropathy (CIPN), a side effect of many anti-cancer drugs including the vinca alkaloids, is characterized by a severe pain syndrome that compromises treatment in many patients. Currently there are no effective treatments for this pain syndrome except for the reduction of anti-cancer drug dose. Existing data support the model that the pain associated with CIPN is the result of anti-cancer drugs augmenting the function of the peripheral sensory nociceptors but the cellular mechanisms underlying the effects of anti-cancer drugs on sensory neuron function are not well described. Studies from animal models have suggested a number of disease etiologies including mitotoxicity, axonal degeneration, immune signaling, and reduced sensory innervations but these outcomes are the result of prolonged treatment paradigms and do not necessarily represent the early formative events associated with CIPN. Here we show that acute exposure to vinca alkaloids results in an immediate pain syndrome in both flies and mice. Furthermore, we demonstrate that exposure of isolated

sensory neurons to vinca alkaloids results in the generation of an inward sodium current capable of depolarizing these neurons to threshold resulting in neuronal firing. These neuronal effects of vinca alkaloids require the transient receptor potential ankyrin-1 (TrpA1) channel, and the hypersensitization to painful stimuli in response to the acute exposure to vinca alkaloids is reduced in TrpA1 mutant flies and mice. These findings demonstrate the direct excitation of sensory neurons by CIPN-causing chemotherapy drugs, and identify TrpA1 as an important target during the pathogenesis of CIPN.

5.1b Results

The sensitivity of the vinca-evoked Na+ currents in C4da neurons to Ruthenium red and HC-030031 suggested the involvement ofdTrpA1 ion channels in the effects of vinca alkaloids on the hyperexcitability of these cells [42,65]. To support the identification of dTrpA1 as the target of vinca alkaloids in C4da neurons we performed RNA-seq analysis on purified C4da neurons. Average gene expression was determined for all Drosophila TRP family channels and revealed that dTrpA1 is the highest expressed TRP channel in C4da neurons (Figure 5.1A) and shows the most enrichment within C4da neurons compared to other tissues (Figure 5.1B). These data are consistent with C4da neurons expressing dTrpA1 channels and support the pharmacological data suggesting that dTrpA1 is the target of vinca alkaloids during the activation of C4da neurons.

5.2 Ion channel expression differences do not underlie Orco mutant defects in sensory responses

The following selection is adapted with minimal modification from:

Utashiro N, Williams CR, Parrish JZ, Emoto K. (2018) Prior activity of olfactory receptor neurons is required for proper sensory processing and behavior in *Drosophila* larvae. Scientific Reports 8:8580.

5.2a Abstract

Animal responses to their environment rely on activation of sensory neurons by external stimuli. In many sensory systems, however, neurons display basal activity prior to the external stimuli. This prior activity is thought to modulate neural functions, yet its impact on animal behavior remains elusive. Here, we reveal a potential role for prior activity in olfactory receptor neurons (ORNs) in shaping larval olfactory behavior. We show that prior activity in larval ORNs is mediated by the olfactory receptor complex (OR complex). Mutations of Orco, an odorant co-receptor required for OR complex function, cause reduced attractive behavior in response to optogenetic activation of ORNs. Calcium imaging reveals that Orco mutant ORNs fully respond to optogenetic stimulation but exhibit altered temporal patterns of neural responses. These findings together suggest a critical role for prior activity in information processing upon ORN activation in Drosophila larvae, which in turn contributes to olfactory behavior control.

5.2b Results

One might imagine that ion channel gene expression in Orco mutant ORNs could account for different patterns of activation following stimulation. To test this, we assayed for changes in ion channel gene expression between wild-type and Orco mutant dorsal organs using RNA-Seq analysis (Figure 5.2a), and found no significant changes in ion channel gene expression in Orco mutant dorsal organs; among 148 ion channel genes queried, only Orco and 4 OR genes exhibited significant differential expression between wild-type and Orco mutant samples (Figure 5.2b). These results suggest that the differences in neural responses in Orco mutant ORNs are unlikely to be caused by changes in ion channel expression.

5.3 The microtubule acetylase dTat is highly expressed in the peripheral nervous system

The following selection is adapted with minimal modification from:

Yan C, Wang F, Peng Y, Williams CR, Jenkins B, Wildonger J, Kim HJ, Perr JB, Vaughan JC, Kern ME, Falvo MR, O'Brien ET 3rd, Superfine R, Tuthill JC, Xiang Y, Rogers SL, Parrish JZ. (2018) Microtubule acetylation is required for mechanosensation in *Drosophila*. Cell Rep 25(4):1051-1065.

5.3a Abstract

At the cellular level, α-tubulin acetylation alters the structure of microtubules to render them mechanically resistant to compressive forces. How this biochemical property of microtubule acetylation relates to mechanosensation remains unknown, although prior studies have shown that microtubule acetylation influences touch perception. Here, we identify the major Drosophila a-tubulin acetylase (dTAT) and show that it plays key roles in several forms of mechanosensation. dTAT is highly expressed in the larval peripheral nervous system (PNS), but it is largely dispensable for neuronal morphogenesis. Mutation of the acetylase gene or the K40 acetylation site in α-tubulin impairs mechanical sensitivity in sensory neurons and behavioral responses to gentle touch, harsh touch, gravity, and vibration stimuli, but not noxious thermal stimulus. Finally, we show that dTAT is required for mechanically induced activation of NOMPC, a microtubule-associated transient receptor potential channel, and functions to maintain integrity of the microtubule cytoskeleton in response to mechanical stimulation.

5.3b Results

To define biological functions of dTAT, we first examined dTat expression patterns. We fluorescence-activated cell sorting (FACS) isolated different GFP-labeled cell types and subjected the cell lysates to RNA-sequencing (RNA-Seq) analysis (Figure 5.3). Among the larval cell types that we surveyed, including muscle, epithelia, glia, and neurons, dTat expression is the highest in neurons. PNS neurons, particularly da neurons that mediate responses to mechanical, thermal, light, and proprioceptive stimuli, highly express isoforms for both dTAT-S and dTAT-L, suggesting that the PNS is likely an important functional site for dTAT.

5.4 Chapter Five Figures

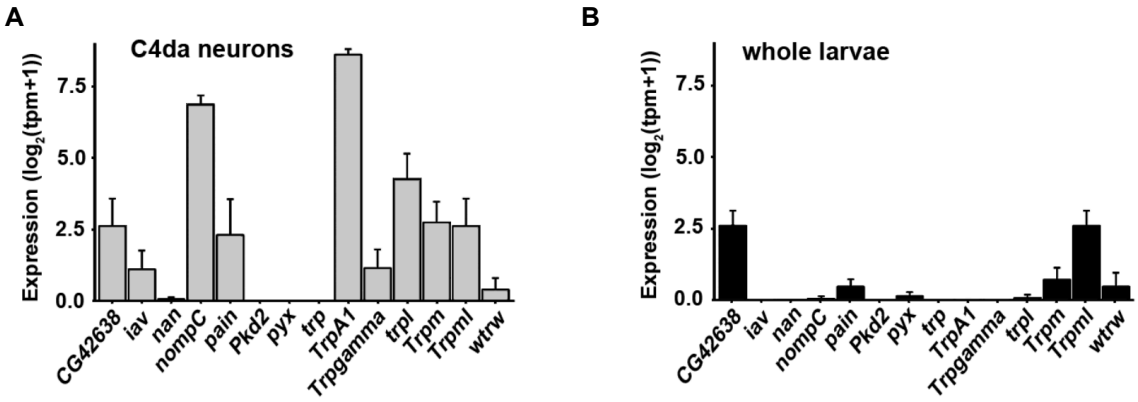

Figure 5.1. Trp channel expression is enriched in neurons.

Graphs represent the average gene expression level of all transcripts of the indicated *Drosophila* TRP family genes in purified C4da neurons (A; gray bars) and in whole larvae (B; black bars). tpm = transcripts per million. Error bars = sem. n = 4 independent samples for each condition.

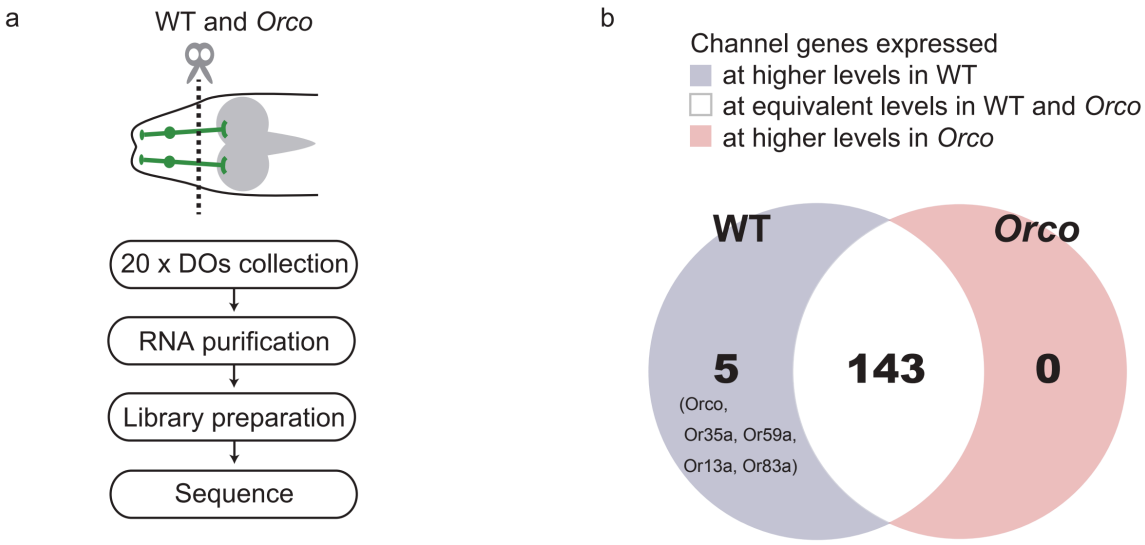

Figure 5.2 RNA-Seq analysis of ion channels in dorsal organs.

(a) Schematic Diagram of RNA-Seq analysis. DOs indicates dorsal organs. (b) Differential expression (DE) analysis of ion channel gene expression in wild-type and Orco mutant dorsal organs. Venn diagram shows the number of ion channel genes that were expressed at significantly higher levels in WT (Blue), at higher levels in Orco mutants (Red), or at equivalent levels in WT and Orco mutant dorsal organs (White). 127 genes were excluded from DE analysis due to insufficient information for expression estimation. Genotype: WT, *w1118;Orco-Gal4,UAS-CD8::GFP/Orco-Gal4,UAS-CD8::GFP*. Orco, *w1118;Orco-Gal4,UAS-CD8::GFP/Orco-Gal4,UAS-CD8::GFP;orco1 /orco1*.

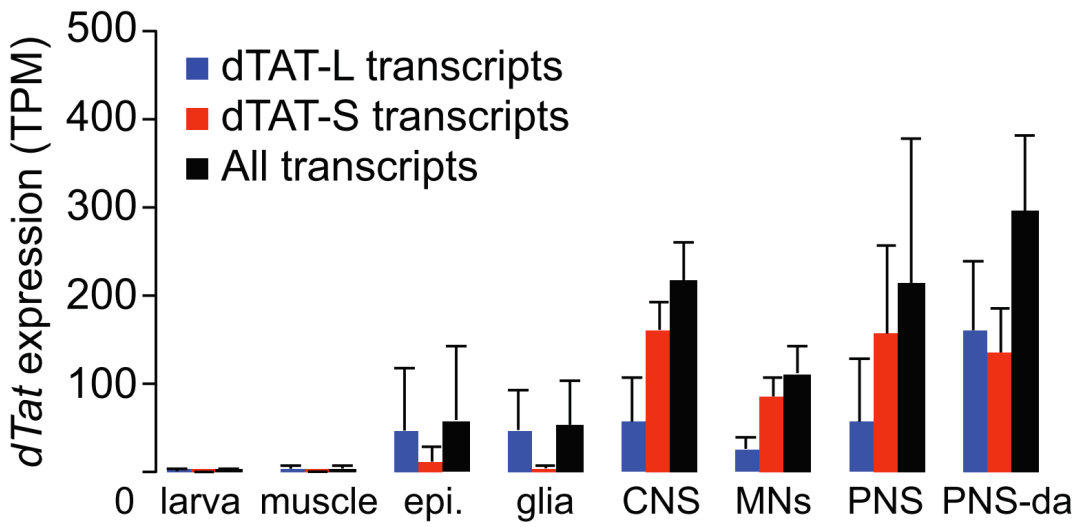

Figure 5.3 RNA-Seq analysis of dtat expression

Bars depict mean expression levels of dTat transcripts in the indicated cell types. TPM, transcripts per million. Error bars, SDs. N = 4+ independent samples for each condition.

Chapter 6 : EPIDERMAL CELLS AUGMENT NOCICEPTION IN *DROSOPHILA MELANOGASTER*

Chapter 6 contains unpublished work resulting from a collaboration between the following scientists: Williams CR*, Yoshino J*, Morita T, Mali S, Emerson CE, Arp CJ, Miller SE, Goltseker K, Prater C, Bautista D, Emoto KE, Parrish JZ

*Authors contributed equally.

6.1 Abstract

Our skin provides a first point of contact with the sensory environment, and many types of cutaneous receptors form specialized terminal structures with epithelial cells that contribute to somatosensation. For example, some low threshold mechanoreceptor afferents form synapse-like contacts with Merkel cells, which directly respond to mechanical stress and tune gentle touch responses. By contrast, much less is known about functional coupling of epidermal cells and nociceptive somatosensory neurons. However, ultrastructural imaging studies demonstrate that keratinocytes ensheath nociceptive C-fibers in a mesaxon-like structure. We found that *Drosophila* larval epidermal cells similarly ensheath nociceptive class IV dendrite arborization (c4da) neurons, that this ensheathment is controlled by an evolutionarily conserved morphogenetic mechanism, and that this ensheathment modulates nociception. We also found that *Drosophila* larval epidermal cells express a variety of sensory channels, as has been reported in vertebrates, suggesting that epidermal cells may actively participate in sensory transduction. We therefore hypothesized that nociception is shaped by sensory-evoked responses of epidermal cells which transduce signals via epidermal sheaths to potentiate nociceptor activation.

To assess whether manipulation of epidermal cells could induce behaviors in *Drosophila melanogaster*, as it can in mice, we used optogenetics. Pan-epidermal optogenetic stimulation elicited robust responses that included stereotyped behaviors associated with activation of nociceptive c4da neurons and mechanosensitive c3da neurons. By calcium imaging of neurons, we found that epidermal depolarization leads to c4da and c3da activation, supporting a model of direct coupling between the epidermis and sensory neurons. To systematically investigate functional coupling of epidermal cells and somatosensory neurons (SSNs) we conducted a targeted optogenetic screen of distinct populations of

body wall epidermal and epidermis-associated cell types. We found that stimulation of different epidermal cell types, but not other body wall cell types, elicited a variety of nociceptive behavioral responses, suggestive of differential coupling of epidermal cells to SSNs. Further, *Drosophila* epidermal cells exhibit sensory-evoked calcium responses in semi-intact fillets and individually exhibit calcium influxes in response to osmotic stress, partially dependent on the mechanosensory channel Piezo. Finally, the epidermal membrane potential is behaviorally relevant. When the epidermis is depolarized, larvae become sensitized to nociceptive stimuli. Overall, we identified a novel sensory role for the epidermis in an invertebrate system, and provided a genetically accessible model for understanding the functional interaction between neurons and epidermis that will better inform our studies of this communication across animals.

6.2 Background

Somatosensation allows us to interpret a wide array of external stimuli so that we may respond quickly with appropriate behaviors. This broad sensory category encompasses tactile, thermal, chemical, proprioceptive, and vibratory impacts and the integration across these different inputs, all in direct contact with our skin, is vital to survival in an ever-changing environment.

Mammalian skin is innervated by multiple types of sensory neurons which can detect all of these stimuli specifically or in combination[1–3]. The skin is predominantly composed of keratinocytes forming a stratified epidermis and these epidermal cells play many important functions including forming a protective barrier against hydration loss, preventing the free entry of harmful microbes, and initiating immune responses[4–6]. Emerging evidence suggests that non-neuronal skin cells and structures, including keratinocytes, Meisner corpuscles, Pacinian corpuscles, keratinocytes, and Merkel cells, additionally contribute to somatosensation. Merkel cells, specialized epidermal cells found at touch domes in regions of high tactile acuity, mediate touch sensation via Piezo2 mechanosensitive channels and release norepinephrine at synapse-like structures onto A-β low-threshold sensory neuron afferents to induce extended and persistent firing[7–9]. Keratinocytes, comprising about 95% of epidermal layers, release inflammatory cytokines to sensitize neurons, express multimodal sensory Trp channels, depolarize in response to mechanical touch and potentiate aversive behaviors via the release of ATP[10–15]. Beyond

epidermal cells, cutaneous Schwann cells also depolarize to harsh touch and can induce nociceptive behaviors when optogenetically stimulated[16]. It is an open question how widely conserved this ability to sense mechanical stimuli and induce behaviors is across peripheral cell types, mammals, and organisms more broadly. Further, basic mechanisms of how epidermis to sensory neuron transduction is functionally relevant are poorly understood.

To investigate the evolutionarily conserved nature of this non-neuronal somatosensation, one approach is to turn to an invertebrate with a simpler set of peripheral cell types and ample genetic tools for cell type specific manipulation. The larval fruit fly *Drosophila melanogaster* has a monolayer of epidermal cells that, in concert with secreted molecules of the cuticle, form an external barrier, coating the entire animal. Similar to their mammalian counterparts, *D. melanogaster* epidermal cells are closely associated with somatosensory neurons and ensheath the dendritic processes of these neurons[17]. There are four classes of dendritic arborization (da) somatosensory neurons in *D. melanogaster* larvae categorized as class I (c1da) through class IV (c4da) in increasing dendritic process complexity[18]. Similar to mammalian sensory neurons, these da neurons have different functions. C1da neurons are proprioceptors[19], c2da and c3da sense cold and gentle touch[20], and c4da are multimodal nociceptors, sensing harsh touch, high temperatures, noxious chemicals, and blue light[21–27]. Of these four, c4da neurons are most closely associated with epidermal cells as measured by the percent of dendritic arbor ensheathed[17], so they are the top candidate for transducing a functional interaction between epidermal cells and somatosensory neurons.

Here, to test the evolutionary conservation of epidermal sensation, we investigated whether non-neuronal peripheral cells contribute to sensation in *D melanogaster*. We found that optogenetic activation of epidermal cells led to nociceptive and non-nociceptive behaviors, mediated by canonical neuronal circuitry. In a broad optogenetic screen of 12 cell types and subtypes in the periphery, we concluded that the induction of nociceptive behaviors is specific to epidermal and neuronal cell types. Epidermal cells express sensory channels and respond directly to mechanical stimuli with calcium influxes, dependent on the presence of the Piezo channel. Finally, the epidermal activation state is behaviorally relevant such that epidermal depolarization can augment larval behavioral responses to natural stimuli. These data are

the first report of epidermal sensation in an invertebrate system, suggesting that epidermal contributions and functionality have been broadly conserved across evolution.

6.3 Methods

Fly husbandry Fly stocks were maintained on standard cornmeal-molasses-agar at 25°C on a 12 hour light - 12 hour dark light cycle. For all experiments involving optogenetic manipulations, larvae were raised in the dark at 25°C on Nutri-Fly Instant Food (Genesee Scientific #66-117), supplemented with 1mM all-trans retinal (ATR; Sigma #R2500) when indicated.

IMAGING

Optogenetic behavior -- high-resolution imaging (Figure 1) One larva at a time was placed on a 100mm petri dish containing 1% agarose. The larva was exposed to 585 nm red light from above by a LED illuminator (SPECTRA X, Lumencor) equipped with a filter (FF01 585/40-25, Semrock) and images were captured with a sCMOS-Camera (Zyla4.2, Andor) at a frame rate of 20 Hz. Larvae were constantly illuminated from above by a far red (940 nm) light source (LDR2-132IR2-940-LA, CSS) for imaging. Larvae were fed (ATR+) or not fed (ATR-) the cofactor ATR as indicated.

Optogenetic behavior -- low-resolution screen (Figure 3) One larva at a time was placed on a 100mm petri dish containing 1% agarose. The larva was exposed to 488 nm green light from above by a CoolLED PE300ultra and videos were captured with a Hamamatsu ORCA Flash CMOS camera at a frame rate of 20 Hz, controlled by MetaMorph. All larvae were fed ATR.

Optogenetic behavior analysis Videos of individual larvae responding to light stimuli were scored on a second-by-second basis using ImageJ[28] or on a frame-by-frame basis using the annotation software BORIS[29]. Behaviors scored, along with descriptions of the criteria for each behavior, are in Table 6.1. Video analysts were blind to the genotype and treatment during scoring. Scoring on a training set was

compared across all analysts to calibrate, and any behaviors for which the primary analyst was unsure were reviewed by an additional analyst.

Calcium imaging ventral nerve cords. Third-instar larvae were dissected along the dorsal midline and pinned on a sylgard-coated dish (Silpot 184, Dow Corning Toray). The internal organs except for neural tissues were removed. Larvae were bathed in HL3.1[30] modified to remove calcium (see Table 6.2 for recipe) in order to minimize larval movement. The ventral nerve cord was imaged using an Olympus BX51WI microscope, equipped with a spinning-disk confocal unit Yokogawa CSU10 (Yokogawa) and an EM-CCD digital camera (Evolve, Photometrics). For activation of epidermal cells with the light gated CsChrimson, red light was delivered by a pE-300 (CoolLED) equipped with a filter (ET645/30x, Chroma). Obtained images were analyzed using Metamorph (https://www.moleculardevices.com/systems/metamorph-research-imaging) and ImageJ[28]. Baseline fluorescence was calculated as the average fluorescence intensity of an ROI over the ten frames prior to light stimulus delivery.

Calcium imaging fillets Third-instar larvae were dissected along the ventral midline and pinned on sylgard with the internal surface facing towards the microscope. All internal organs, including the central nervous system, were removed. Larvae were bathed in calcium-containing (i.e. original) HL3.1[30] (see Table 6.2 for recipe) except where indicated and images of the dorsal midline between abdominal segments A2 and A4 were captured with a Zeiss Axio Zoom V16 microscope. Obtained images were analyzed using ImageJ[28]. Due to increased bleaching observed with the Axio Zoom setup, bleaching rates were determined for each cell over the first baseline minute of each movie by fitting an exponential decay curve to the fluorescent intensity, and fluorescent intensities across each movie were corrected using this empirically determined bleaching rate. *Mechanical stimulus*: Fillets were poked with a tapered borosilicate capillary with a rounded tip, using a micromanipulator to induce a deflection of 25μm. The cell directly contacted by the capillary was analyzed, as was a second cell in the vicinity but not contacted to serve as a negative control. *Thermal stimulus*: Fillets were pinned on top of a Peltier heat plate and temperature was monitored with a thermocouple placed in the HL3.1 drop adjacent to the fillet. After

baseline recording, the plate was turned on and the temperature of the bath was raised to 43°C, which took an average of 30 seconds, and was held there for an additional 30 seconds. Ten epidermal cells per fillet were analyzed. *Chemical stimulus*: Fillets were bathed in mineral oil and after baseline recording allyl isothiocyanate (AITC; Sigma 377430) was added to a final concentration of 100mM. Ten epidermal cells per fillet were analyzed. *Vibration stimulus*: Fillets were placed on top of a speaker, and the speaker was used to deliver a 5-second 70dB, 500Hz pure tone. Ten epidermal cells per fillet were analyzed.

Calcium imaging single cells. Three to eight larval fillets were dissociated in 400µL of 50% PBS / 50% Schneider's media with 200U/mL collagenase type I (Fisher 17-100-017), with mixing at 1000RPM at 33°C for 12 minutes, with trituration every 6 minutes. Undigested fillets were removed and the remaining suspension was spun at 500g for 3 min, followed by aspiration of the supernatant down to a 25 µL cell suspension. Cells were resuspended in 150 µL fresh PBS / Schneider's solution and plated onto poly-D-lysine (1mg/ml, Sigma P7886) coated No.1 coverslips, with 25 µL cell solution per coverslip. Cells were cultured at least 30 min and up to 3 hrs at 25°C prior to imaging. Cells were imaged using a Zeiss Observer inverted microscope with a 20x objective at a frame rate of 0.33 Hz. Solutions (see Table 6.2 for recipes) were perfused in using a perfusion system. At the end of each imaging session, 1µM ionomycin was perfused in and only cells that showed a calcium response, as defined by a 10% increase from baseline fluorescence, were used in analysis. Obtained images were analyzed using Metafluor and IgorPro and baseline fluorescence was calculated as the average fluorescent intensity of an ROI over 10 frames prior to stimulus delivery. A hypotonic calcium response was defined as a 10% or greater increase from baseline fluorescence.

Morphology imaging For peripheral imaging of cellular morphology, live single larvae were mounted in 90% glycerol under a coverslip and imaged on a Leica SP5 confocal microscope using a 40x 1.25NA lens. For central imaging, larvae were dissected on sylgard plates, briefly fixed in 4% paraformaldehyde (PFA) in PBS for 15 minutes at room temperature, washed 3 x 5 min in PBS, and mounted for imaging.

TRANSCRIPTOMIC ANALYSES

RNA isolation for RNA-Seq Larvae with cytoplasmic GFP expressed in different target tissue types were microdissected and dissociated in 400µL of 50% PBS / 50% Schneider's media with 200U/mL collagenase type I (Fisher 17-100-017) and 1% BSA. Dissociating solutions were mixed at 1000RPM at 37°C for 18 minutes, with trituration every 6 minutes. After dissociation, cells were transferred to a new 35mm petri dish with 1mL 50% Schneider's media, 50% PBS supplemented with 1% BSA. Under a fluorescent stereoscope, individual fluorescent cells were manually aspirated with a glass pipette into PBS with 0.5% BSA, and then serially transferred until isolated without any additional cellular debris present. Ten cells per sample were aspirated together and transferred to a mini-well containing 3ul lysis solution (0.2% Triton X in water with 2U / ul RNAse Inhibitor). After pipetting up and down to quickly lyse the cells, the lysed cell solution was transferred to a microtube and stored at -80C until library preparation. For the picked cells, 2.3ul of lysis solution was used as input for library preparation. For two cell types, muscles and trachea, we were unable to successfully dissociate single cells from the tissue. Instead, entire tissues were dissected and snap frozen in RNAqueous lysis buffer (Thermofisher). RNA was isolated using the RNAqueous Micro Kit (Thermofisher #AM1931) and diluted down to 0.2 ng/uL to use as input for RNA-Seq library preparation.

RNA-Seq library preparation RNA-Seq libraries were prepared from the picked cells and the tissue-isolated RNA following the Smart-Seq2 protocol for full-length transcriptomes[31]. For both, libraries were started with 2.3ul of lysis or RNA. To minimize batch effects, primers, enzymes, and buffers were all used from the same lots for all libraries. Libraries were multiplexed, pooled, and purified using AMPure XP beads, quality was checked on an Agilent TapeStation, and libraries were sequenced as 51-bp single end reads on a HiSeq4000 at the UCSF Center for Advanced Technology, at read depths ranging from 0.8 million to 7 million reads.

RNA-Seq data analysis Reads were demultiplexed with CASAVA (Illumina) and read quality was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC[32]. Reads containing adapters were removed using Cutadapt version 2.4[33] and reads were mapped to the *D.*

melanogaster transcriptome, FlyBase genome release 6.29, using Kallisto version 0.46.0[34] with default parameters. Two samples were removed from further analysis for poor quality, defined as low read depth (<0.8 million reads) and low mapping rate (< 45%). t-SNE analysis was implemented using the R package Rtsne[35], using the 500 most variable genes as input.

BEHAVIORAL ASSAYS

Von Frey assay Larvae were placed on a scratched 35mm petri dish with a thin sheet of water such that larvae stayed moist but did not float. A Von Frey filament, made by adhering fishing line to a glass capillary and calibrated to deliver an ~20mN or ~49mN force upon bending, was poked into the dorsal side of the larva between segments A2 and A6 until the filament bent and held for two seconds[36]. Upon stimulus removal, the number of nocifensive rolls exhibited by the larva within ten seconds was recorded. Roll count data were fit with a Poisson regression model to assess differences between ATR+ and ATR- conditions and were corrected for multiple comparisons with a Benjamini-Hochberg correction.

6.4 Results

6.4a Optogenetic activation of epidermal cells induces nociceptive behaviors in D. melanogaster

Skin cells in mammals can induce and augment avoidance behaviors when stimulated[7,12], yet it is unknown whether this capability extends to other evolutionary groups, such as arthropods. To test if epidermal stimulation could similarly elicit behaviors in an invertebrate, we used optogenetics to stimulate the epidermis of *D. melanogaster* larvae. The light-gated channel CsChrimson is a variant of channelrhodopsin that has a red-shifted activation curve, allowing for red light stimulation rather than blue light stimulation[37], which is ideal since blue light lowly activates nociceptive sensory neurons in *D. melanogaster*[38]. As other channelrhodopsins do, CsChrimson requires the small molecule all-trans-retinal (ATR) to open in response to light, and this can be supplemented in the food provided to larvae since *D. melanogaster* do not naturally synthesize ATR[39]. CsChrimson is a nonselective cation channel[37] that likely causes both depolarization and an increase in internal calcium. As we do not know which of these

two potential signaling mechanisms is the key component of CsChrimson-induced responses in any stimulated cell, in this paper we will refer to optogenetically stimulating cells as activating the cells.

We targeted CsChrimson expression broadly to the epidermis using the Gal4-UAS system, with *R38F11-Gal4* and *UAS-CsChrimson-mVenus*. In combination with *elav-Gal80* to suppress expression in a single repeating central neuron, the resulting expression pattern is highly specific to body wall epidermal cells (Figure 6.1b; Supplementary Figure 6.6 S1). In third instar larvae, optogenetic activation of epidermal cells induced robust and prolonged behaviors that lasted throughout the duration of the light stimulus, dependent on the presence of the obligate cofactor ATR (Figure 6.1c,d). These responses were composed of behaviors previously associated with the activation of nociceptive C4da somatosensory neurons, including rolling, rearing, and c-bending[36], and of mechanosensory C3da somatosensory neurons, including hunching, backing, freezing, and pausing[40]. A complete description of the criteria used to define each behavior is in Table 6.1. These epidermally-induced behaviors tended to transition from a set primarily composed of intense C4da-type behaviors (rolling and c-bending) in the first few seconds of stimulation to a mixture of less intense C4da-type behaviors (ie tilting) and C3da-type behaviors (freezing and backing) later during stimulation.

Direct stimulation of somatosensory neurons elicited canonical behaviors in a short burst immediately following activation (Figure 6.1e-g). However, following this initial response, most larvae returned to non-responsive crawling behaviors within five to ten seconds, despite the light stimulus persisting for thirty seconds. This suggests that there was adaptation in the response, either at the somatosensory neurons or further downstream in the neuronal circuits, and that this adaptation did not occur as epidermal stimulation continued. This difference in the response between epidermal stimulation and neuronal stimulation is reminiscent of findings with Merkel cells. When a mechanical stimulus is delivered to mouse skin, somatosensory Merkel cell afferents in isolation fire action potentials initially and then adapt and return to baseline whereas when Merkel cells are present, these afferents continue firing action potentials at a slightly reduced rate after the initial burst, for the duration of the mechanical stimulus[8].

Given that epidermally-induced behaviors mirrored those observed following sensory neuron stimulation, we wondered about the neural circuitry underlying epidermally-induced behaviors, if any; are somatosensory neurons activated downstream of epidermal stimulation?

6.4b Epidermal activation leads to somatosensory neuron depolarization

The most prominent behaviors observed immediately following epidermal activation were c-bending and rolling, which are classically associated with C4da activation[36]. Additionally, previous work has found that C4da neurons are tightly associated with epidermal cells and of the somatosensory subtypes, they are the most highly ensheathed by epidermal cells[17]. Therefore, we began by testing whether C4da neurons were activated downstream of epidermal stimulation using the genetically encoded calcium indicator GCaMP6s in C4da neurons in combination with epidermally-expressed CsChrimson. To avoid interference from the mVenus fluorescent tag on CsChrimson, we imaged the GCaMP signal in C4da axon terminals, located in the ventral nerve cord (VNC) (Figure 6.2a).

In response to epidermal activation, there is a sharp increase in C4da calcium levels with an average maximum fluorescence change of GCaMP6s of 235% within 3 seconds of the onset of epidermal stimulation (Figure 6.2b). This increase is not a cell-autonomous response to the light stimulus since there is no increase in fluorescence in response to the light in the absence of the CsChrimson obligate cofactor ATR. Over the course of the one-minute light stimulus, the calcium levels in C4da decreased but did not return to baseline; the average fluorescence was 27% over baseline at the end of the stimulus, indicating that the C4da neurons were still active and potentially contributing to behaviors. Initial bursting of C4da in response to a stimulus onset followed by reduced activity as the stimulus persists is consistent with responses observed when C4da is activated directly by mechanical or thermal stimuli[42].

In addition to C4da-type behaviors, we also observed C3da-type behaviors following epidermal activation, and the epidermis ensheathes C3da dendrites, though at a lower rate than C4da dendrites. We therefore repeated these experiments with expression of GCaMP6s in C3da neurons (Figure 6.2c). Similar to C4da neurons, there is an increase in C3da axonal calcium levels following epidermal activation, with an average maximum fluorescent change of 116% (Figure 6.2d). In contrast to C4da

neurons, the average fluorescence returned to baseline levels within ten seconds of the stimulus onset. This adaptation has also been observed when C3da neurons are stimulated by a persistent gentle touch[43]. Thus, C3da neurons were activated to a lesser extent and for a shorter duration following the same epidermal stimulation than C4da neurons.

One other contrast between the C4da and the C3da calcium profiles was a spatial difference. Both C4da and C3da neurons are found in every segment of the larva, and our light stimulation was not localized, so the epidermis was activated in every segment. In the VNC, the largest responses of C4da axons were observed in the posterior-most segments, abdominal segments A6 - A8, whereas calcium transients were more widely distributed across all abdominal segments for C3da imaging. For consistency of quantification, we manually chose ROIs within segments that showed a light response when possible. Within a given preparation, the same segment responded repeatedly to repeated stimulation (data not shown), suggesting there is specificity in the responding neurons. At this point, we are unable to conclude whether the posterior-localized C4da response or the distributed C3da responses are technical artifacts. If they are not, this provokes speculation that somatosensory neurons in different segments may be differentially responsive to epidermal depolarization, perhaps due to differential coupling of the epidermis and neurons in different segments. This raises the question of whether all peripheral epidermal cells are similarly able to activate sensory neurons, or if there is some spatial or subtype specificity in the behaviors evoked by epidermal cells.

6.4c Nociceptive behaviors are specific to epidermal activation among peripheral non-neuronal cells

To address this question of specificity of the behavioral responses induced by different cell types, we began by gathering an atlas of cell-type specific Gal4 drivers to independently control 12 different cell types in the larval body wall. These included four drivers with restricted expression in the epidermis, ranging from broadly expressed (eg. *R38F11-Gal4*) to striped expression patterns (eg. *hh-Gal4*) to defined subtypes (eg. apodemes, *sr-Gal4*) (Figure 6.3a); four drivers for somatosensory neurons, either individual da neurons or a pan-da driver (Figure 6.3b); and four drivers for other cell types including muscles, hemocytes, oenocytes, and trachea (Figure 6.3c). For any drivers that showed additional, non-

target expression in neurons, we also included *elav-Gal80* in the optogenetic experiments to suppress neuronal expression.

We next performed a comprehensive optogenetic screen of these peripheral cell types by expressing CsChrimson under the control of each cell-type specific driver and recording the response to a light stimulus. For technical reasons, we used a green light stimulus (488 nm) in this screen, which provides approximately 25% maximal activation of CsChrimson[37], and delivered a 10 second light stimulus followed by 15 seconds of recording without stimulation. We observed a similar range of behaviors as previously described with epidermal, C4da, or C3da stimulation, and further observed instances of peristalsis defects (Figure 6.3d,e). Consistent with our previous findings with the broad epidermal driver *R38F11-Gal4*, all of the various epidermal drivers induced immediate and short duration nociceptive, C4da-type behaviors, with an average duration of 5 seconds. However, after this initial response many larvae returned to non-responsive crawling for the remainder of the stimulus. This is potentially due to the lower intensity light stimulus used for this broader screen.

Aside from epidermal activation, activation of other cell types also induced behaviors. Surprisingly, C1da activation led to rearing and hunching behaviors, which contrasts the peristalsis defects expected based on previous reports of c1da-relevant behaviors[19]. C2da activation led to rolling and c-bending as expected[44] and pan-da activation led primarily to long-lasting C4da-type behaviors followed by peristalsis defects, which is consistent with previous reports that coincident stimulation of multiple neuron types augments nociceptive responses[45]. Neuronal stimulation induced non-crawling behaviors throughout the stimulation, with behaviors often persisting after the stimulus was removed. One potential explanation for the difference between this duration and that observed in Figure 6.1 is that this was a lower intensity stimulation, so perhaps it took a longer period of time for neurons to adapt to this stimulus. Further, c4da neurons were likely lowly stimulated throughout the light stimulus due to low intensity of blue light in the stimulus. Most of the other cell types induced no changes in behavior. The one exception among this group of other cell types was muscles, activation of which induced hunching throughout the stimulation, followed by persistent freezing that continued at least 15 seconds after the stimulus ended. Our demonstration that nociceptive behaviors are not broadly induced by all non-neuronal cell types in the body wall but instead are specific to epidermal cells suggests that epidermal

cells are uniquely poised to contribute to sensation. Therefore, we next asked when epidermal cells might be stimulated endogenously.

6.4d Epidermal cells are intrinsically mechanically sensitive

As a first step to understand what stimuli may naturally activate epidermal cells and to understand how epidermal cells may be specifically tuned in ways that other non-neuronal cells are not, we profiled the transcriptomes of all the cells in our atlas of body wall cell types. For each of the 11 cell types for which we had a specific driver, we collected at least ten 10-cell samples from larval segments A2 - A4 to generate RNA-Sequencing (RNA-Seq) libraries. The pan-da driver *21-7-Gal4* was excluded as we had more specific drivers for all neuron types that it labeled. Because these samples were all collected in the same manner and prepared in the same way, this is a rich dataset to use for interrogating similarities and differences between cell types. A t-SNE plot, which serves as a visualization of the local and global relationships between these transcriptomes, shows broad structure as would be expected (Figure 6.4a). The different types of epidermal cells are similar in their transcriptomes, neuronal profiles look broadly similar to one another, and the other cell types are more distinct. Within these broad cell type groups, there is some structure, but it is only the 'other' cell types that clearly separate out by type.

Beyond this overview of cell type similarity, the transcriptomes can be probed to examine specific genes. To understand how non-neuronal cells may naturally respond to stimuli, we looked at the expression of multiple known sensory channels (Figure 6.4b). These were genes that were previously identified in somatosensory neurons as contributing to mechanosensation (*Piezo*, *painless*, *TrpA1*, *nompC, ppk*)[21,22,24,25,43], vibration sensation (*nompC*, *nan*, *iav*)[46], or thermosensation (*Trpa1, pyrexia*, *subdued*)[47–49]. All of these channels were most highly expressed in neurons, but also showed expression elsewhere. These results are also in agreement with previously reported expression patterns. For example, the multimodal channel *TrpA1* is most highly expressed in C4da neurons[25]. Three channels, *subdued*, *pyrexia*, and *Piezo*, were broadly expressed, albeit at lower levels than in neuronal samples, across cell types, including many of the epidermal cell types. This expression pattern supports a model in which epidermal cells could respond to thermal or mechanical stimuli.

One way to test this model of epidermal responsiveness is to deliver a stimulus directly to epidermal cells and monitor responses. As calcium is a common cell-signaling molecule, we chose to use GCaMP6s, expressed in epidermal cells, to monitor responses to various stimuli as an initial screen of response properties. We delivered mechanical, thermal, chemical, or vibration stimuli to larvae in semi-intact fillet preparations and quantified the resulting change in GCaMP fluorescence in epidermal cells (Figure 6.4c, d). The intensity of the stimuli was chosen to be strong enough to induce GCaMP fluorescence increases in neurons known to respond to each of these stimuli (i.e. chorodotonal neurons for the vibration stimulus and C4da neurons for all others). Epidermal cells were most responsive to mechanical stimuli, with the greatest change in fluorescence reaching over 200%, though about half of the epidermal cells probed showed little to no increase following mechanical stimulation, suggesting substantial variability in the responsiveness of different epidermal cells. As epidermal cells are the first point of contact for external mechanical stimuli, this is likely to be a natural or frequent stimulus for epidermal cells to encounter. The epidermal cells were also responsive to high temperatures with high variability, showing both increases and decreases in fluorescence up to about 80%. There was no epidermal change in fluorescence in response to vibration or chemical stimuli, which is consistent with the absence of *nan*, *iav*, and *TrpA1* expression in the epidermis. During this screen, we observed that the GCaMP fluorescence levels stayed elevated in epidermal cells longer after activating stimuli than it did in neurons, suggesting a potentially distinct mechanism of homeostatic monitoring of calcium levels. Due to issues with bleaching, we were unable to quantify more specifically this difference in calcium perdurance in this imaging setup.

To investigate whether this observed mechanical responsiveness was an emergent property of epidermal cells in a sheet or was intrinsic to the cells, we chose to test mechanical responsiveness in isolated cells. We dissociated epidermal cells expressing GCaMP6s from the body wall and then delivered stimuli directly to single cells. Altering the osmotic pressure on a cell can induce cell swelling and cause mechanical stress, so has been used previously as a model for mechanical stimuli. When epidermal cells were subjected to a hypotonic solution, predicted to cause cell swelling, a portion of the cells showed an immediate spike in internal calcium levels (Figure 6.4e, f, g). The hypotonic solution was followed with ionomycin to confirm cell viability. Among the viable cells, 44% of epidermal cells responded

to osmotic stretch, defined as an increase of 10% or more in GCaMP fluorescence. This suggests both that epidermal cells are mechanosensitive and that there is variability in this responsiveness, as previously observed with our semi-intact fillet GCaMP screen. The top candidate among known mechanosensory channels to explain this responsiveness is Piezo, since this gene was broadly expressed in the epidermis, potentially conferring mechanosensitivity. When we repeated the osmotic stress GCaMP imaging in a *piezo* mutant background, we observed a decrease in the percent of epidermal cells responding, down to 19% (Figure 6.4h). Among responding cells, the fluorescence increase was smaller, suggesting that mechanical responsiveness of epidermal cells is muted in the absence of *piezo*. Because the osmotic stretch response was not fully eliminated in the *piezo* mutant, there are likely additional contributions from other sensory channels. Together, these experiments demonstrate that, both in a sheet and as individual cells, epidermal cells are responsive to mechanical stimuli and, further, this response is at least partially dependent on *piezo*.

6.4e Epidermal activation state is behaviorally relevant

Given that epidermal cells are mechanosensitive and that their activation can lead to nociceptive behaviors, we next asked whether the activation state of epidermal cells influenced the behavioral response to nociceptive stimuli, such as a mechanical harsh touch. Exposure of larvae expressing the optogenetic channel CsChrimson in their epidermal cells to ambient broad spectrum light under a dissecting microscope, 30 +/- 3 mW / cm^2, is insufficient to induce the nociceptive behaviors observed when larvae are exposed directly to high intensity red light. Therefore, this baseline ambient light can be considered a subthreshold stimulation of epidermal cells. We tested whether this subthreshold epidermal stimulation augmented larval responses to mechanical stimuli.

When larvae are poked with a calibrated Von Frey filament, they perform characteristic nociceptive rolls to escape the stimulus[36]. This behavior requires C4da neurons as well as multiple described interneurons downstream in the escape circuitry[36,44,45,50–52]. We exposed larvae expressing CsChrimson in the epidermis, as well as relevant genetic control genotypes, to a harsh touch with a 49mN von Frey filament or a low threshold touch with a 20mN von Frey filament and counted the number of nociceptive rolls performed in the presence or absence of the CsChrimson obligate cofactor ATR

(Figure 6.5). The key comparison to make is between ATR+ and ATR- for a given genotype, since there may be differences in baseline responsiveness between genetic backgrounds. Larvae were more responsive with ATR, as defined by a higher probability of performing more nociceptive rolls, when elicited by either the harsh touch or low threshold stimulus ($p=8.1x10^{-3}$ and $p=4.2x10^{-5}$, respectively). There was a minor increase in roll numbers with one genetic control in response to the low threshold stimulus, but this increase was of much smaller magnitude than the effect of epidermal CsChrimson expression. Overall, our studies with natural stimuli support a role for the epidermis in responding to external inputs.

6.5 Discussion

Epidermal cells broadly cover organisms' exteriors, yet the potential sensory functions of these cells have only been investigated in a small number of mammals. Here, we find that this sensitivity extends beyond vertebrates: epidermal cells of *D. melanogaster* respond directly to mechanical stimuli to augment nociceptive behavioral responses. Specific optogenetic stimulation of epidermal cells induces nociceptive and non-nociceptive behaviors. By combining epidermal optogenetic stimulation with somatosensory neuron calcium imaging, we find that the sensory neurons C4da and C3da are activated downstream of epidermal activation, suggesting that these behaviors are mediated via canonical neuronal circuitry. In a broad optogenetic screen, we observe that many subtypes of epidermal cells can induce nociceptive behaviors whereas other non-neuronal peripheral cells do not induce behaviors, suggesting specificity to the epidermis. Next, we identify the expression of known sensory channels in epidermal cells and find that these cells are responsive to mechanical stimuli, partially dependent on the presence of the conserved mechanosensitive channel Piezo. Finally, with behavioral tests using natural stimuli we demonstrate that the activation state of epidermal cells is functionally relevant, as activating epidermal cells sensitizes larvae to mechanical stimuli.

6.5a Epidermal cells play a conserved role in somatosensation

When did epidermal cells first evolve sensory functions? The contribution of non-neuronal sensation to behavioral responses was previously reported in mammals[8,12,16] and we here report this functionality is also present in an invertebrate. There are multiple scenarios in which sensation or detection of a stimulus must travel from an epidermal surface to the brain. There are broadly two mechanisms by which this transduction may occur. First, stimuli can be directly detected by sensory neurons that project to the brain. Examples of stimuli sensed by this direct transduction mechanism are odorants that bind to olfactory neurons in the vomeronasal cavity of mice[53] or chemical irritants that bind directly to amphid neurons in the amphid sensory neuron of *C. elegans*[54]. In the second mechanism, stimuli are detected by sensory epithelial cells that release a neuromodulator onto sensory afferents that project to the brain. Examples of this latter category include hair cells of the mammalian inner ear and zebrafish lateral line[55,56], gustatory cells of mammalian taste buds[57], and Merkel cells of the mammalian skin[8]. In invertebrate systems, all sensation described to date to our knowledge falls into the former category; our study here is the first to present evidence in an invertebrate of the latter category, indirect sensation. This suggests that the use of intermediary transducing epithelial cells is much more broadly conserved across evolution than previously thought.

6.5b Mechanisms of epidermis - somatosensory neuron coupling

One remaining open question in our work is a functional mechanism explaining how epidermal activation leads to somatosensory neuron depolarization. Epidermal cells ensheath the dendrites of somatosensory neurons and this close association could provide the proximity and specificity needed for functional coupling. In particular, c4da neurons are the most highly ensheathed of all the somatosensory neurons, and so this enrichment may explain the predominance of c4da-type behaviors induced following epidermal activation. There are two models that we propose could underlie this functional coupling. First, epidermal cells could release a neuromodulator onto somatosensory neurons that induces depolarization. This mechanism is akin to that observed for keratoinoctytes and Merkel cells in mice[8,12]. Supporting this model, our transcriptomic data show expression of vesicular release machinery in epidermal cells, including synaptobrevin, Rab11, Scamp, and alphaSNAP. However, we have been unable to identify

secretory vesicles poised to release neuromodulator in the epidermis in electron microscopy imaging we have performed (data not shown). Second, epidermal cells could form gap junctions with somatosensory neurons and thereby directly increase calcium concentrations or depolarization in somatosensory neurons. Supporting this model, our transcriptomic data show the expression of the same set of gap junction genes in epidermal cells, C4da neurons, and C3da neurons, including Innexin-2, which is an obligate component in *in vitro* studies of homomeric and heteromeric gap junctions[58]. Further, gap junction proteins localize to regions close to sites of ensheathment (data not shown), though we have been unable to resolve whether they form channels across epidermal and neuronal cells or are hemichannels and / or intracellular junctions. Ongoing studies using RNAi-mediated knockdown, mutants, and pharmacological agents are addressing leads based on our transcriptomic expression data to test both of these models.

6.5c Calcium signaling in the skin

In this paper, we primarily defined activation of the epidermis as an increase in the calcium concentration. Calcium is extensively used as a signaling molecule, and is likely induced in response to a wide range of stimuli, both external sensory stimuli as described here and internal. In development, epidermal calcium waves are required for proper patterning[59]. In response to cellular damage or wounding in mammals, plants, and flies, calcium levels increase to induce repair mechanisms[60,61]. In *D. melanogaster* larvae, natural mechanical stimuli are likely to also induce wounding, which could amplify the epidermal response. It is possible that the increased sensitivity we observe when activating epidermal cells could be partially mimicking a damage response, in addition to the sensory mechanical response. However, this still supports our model that epidermal activation augments nociception.

6.5d Prolonged nature of epidermal responses

We observed that epidermally-induced responses perdured for much longer than neuronally-induced responses (Figure 6.1). This is similar to the differences in sensory afferent activation in response to a prolonged mechanical stimulus in mice, which is brief in the absence of Merkel cells but

lasts throughout the duration of stimulation when Merkel cells are present[8]. One explanation is that somatosensory neurons may adapt more quickly to a stimulus than epidermal cells do due to the presence of a refractory period following action potential firing. However, the epidermally-induced behaviors were dependent on the activity of C4da neurons (data not shown). Intriguingly, we observed prolonged low level calcium elevation in C4da neurons but not C3da neuron during epidermal stimulation (Figure 6.2), suggesting that this prolonged behavioral response is not due to continuous high intensity activation of all somatosensory neurons. How could behaviors persist if the somatosensory neurons do not remain highly activated? There could be downstream neurons that do not adapt and stay activated following epidermal activation. This would require that epidermally-induced somatosensory neuron activation can be distinguished from neuronally-induced somatosensory neuron activation in downstream neuronal circuitry. We did not attempt to monitor calcium levels in known interneurons downstream in the nociceptive circuitry, but this could be performed in follow up experiments. Alternatively, this prolonged behavior could be a function of the full-body epidermal stimulation that we delivered. While a single C4da neuron does not stay activated through the epidermal stimulation, perhaps a sufficient number of C4da neurons throughout the body are activated downstream of the epidermis at any one time to allow the larva to display a continuous nociceptive behavioral response.

Another observation that we made in regards to differences between epidermal and neuronal stimulation was that calcium levels tended to stay elevated in epidermal cells longer than in neurons. These slower calcium dynamics may be explained by a difference in the homeostatic responses of these two cell types, and could also contribute to prolonged behaviors and sensitization. For instance, if epidermal calcium is elevated by a mechanical stimulus and then the animal receives a second stimulus, the larva may be able to mount a larger or swifter response due to the epidermal priming or 'memory' from the previous stimulus. We are testing this epidermal priming model by delivering multiple stimuli and asking whether epidermal activation state affects the ability to respond to subsequent stimuli.

In conclusion, we show here that epidermal cells augment nociception in *D. melanogaster* and demonstrate more broadly that non-neuronal cells are important targets for fully understanding our interpretation of the outside world.

6.6 References

1. Li, C-L. *et al.* Somatosensory neuron types identified by high-coverage single-cell RNA-sequencing and functional heterogeneity. *Cell Res.* **26**, 83–102 (2016).

2. Koch, S. C., Acton, D. & Goulding, M. Spinal Circuits for Touch, Pain, and Itch. *Annu. Rev. Physiol.* **80**, 189–217 (2018).

3. de Moraes, E. R., Kushmerick, C. & Naves, L. A. Morphological and functional diversity of first-order somatosensory neurons. *Biophys. Rev.* **9**, 847–856 (2017).

4. Taffoni, C. & Pujol, N. Mechanisms of innate immunity in C. elegans epidermis. *Tissue Barriers* **3**, e1078432 (2015).

5. Elias, P. M. Skin barrier function. *Curr. Allergy Asthma Rep.* **8**, 299–305 (2008).

6. Eyerich, K. & Eyerich, S. Immune response patterns in non-communicable inflammatory skin diseases. *J. Eur. Acad. Dermatol. Venereol.* **32**, 692–703 (2018).

7. Maksimovic, S. *et al.* Epidermal Merkel cells are mechanosensory cells that tune mammalian touch receptors. *Nature* **509**, 617–621 (2014).

8. Hoffman, B. U. *et al.* Merkel Cells Activate Sensory Neural Pathways through Adrenergic Synapses. *Neuron* **100**, 1401–1413.e6 (2018).

9. Abraham, J. & Mathew, S. Merkel Cells: A Collective Review of Current Concepts. *Int. J. Appl. Basic Med. Res.* **9**, 9–13 (2019).

10. Koizumi, S. *et al.* Ca2+ waves in keratinocytes are transmitted to sensory neurons: the involvement of extracellular ATP and P2Y2 receptor activation. *Biochem. J.* **380**, 329–338 (2004).

11. Tsutsumi, M. *et al.* Mechanical-stimulation-evoked calcium waves in proliferating and differentiated human keratinocytes. *Cell Tissue Res.* **338**, 99–106 (2009).

12. Moehring, F. *et al.* Keratinocytes mediate innocuous and noxious touch via ATP-P2X4 signaling. *eLife* **7**, (2018).

13. Fuchs, E. Keratins and the Skin. *Annu. Rev. Cell Dev. Biol.* **11**, 123–154 (1995).

14. Li, W.-W., Guo, T.-Z., Li, X., Kingery, W. S. & Clark, J. D. Fracture induces keratinocyte activation, proliferation, and expression of pro-nociceptive inflammatory mediators. *Pain* **151**, 843–852 (2010).

15. Ho, J.-C. & Lee, C.-H. TRP channels in skin: from physiological implications to clinical significances. *Biophysics* **11**, 17–24 (2015).

16. Abdo, H. *et al.* Specialized cutaneous Schwann cells initiate pain sensation. *Science* **365**, 695–699 (2019).

17. Jiang, N. *et al.* A conserved morphogenetic mechanism for epidermal ensheathment of nociceptive sensory neurites. *eLife* **8**, (2019).

18. Grueber, W. B., Jan, L. Y. & Jan, Y. N. Different levels of the homeodomain protein cut regulate distinct dendrite branching patterns of Drosophila multidendritic neurons. *Cell* **112**, 805–818 (2003).

19. Hughes, C. L. & Thomas, J. B. A sensory feedback circuit coordinates muscle activity in Drosophila. *Mol. Cell. Neurosci.* **35**, 383–396 (2007).

20. Tsubouchi, A., Caldwell, J. C. & Tracey, W. D. Dendritic filopodia, Ripped Pocket, NOMPC, and NMDARs contribute to the sense of touch in Drosophila larvae. *Curr. Biol. CB* **22**, 2124–2134 (2012).

21. Tracey, W. D., Wilson, R. I., Laurent, G. & Benzer, S. painless, a Drosophila gene essential for nociception. *Cell* **113**, 261–273 (2003).

22. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. The role of Drosophila Piezo in mechanical nociception. *Nature* **483**, 209–212 (2012).

23. Guo, Y., Wang, Y., Wang, Q. & Wang, Z. The role of PPK26 in Drosophila larval mechanical nociception. *Cell Rep.* **9**, 1183–1190 (2014).

24. Zhong, L., Hwang, R. Y. & Tracey, W. D. Pickpocket is a DEG/ENaC protein required for mechanical nociception in Drosophila larvae. *Curr. Biol. CB* **20**, 429–434 (2010).

25. Zhong, L. *et al.* Thermosensory and nonthermosensory isoforms of Drosophila melanogaster TRPA1 reveal heat-sensor domains of a thermoTRP Channel. *Cell Rep.* **1**, 43–55 (2012).

26. Gorczyca, D. A. *et al.* Identification of Ppk26, a DEG/ENaC Channel Functioning with Ppk1 in a Mutually Dependent Manner to Guide Locomotion Behavior in Drosophila. *Cell Rep.* **9**, 1446–1458 (2014).

27. Neely, G. G. *et al.* A genome-wide Drosophila screen for heat nociception identifies α2δ3 as an evolutionarily conserved pain gene. *Cell* **143**, 628–638 (2010).

28. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

29. Friard, O. & Gamba, M. BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods Ecol. Evol.* **7**, 1325–1330 (2016).

30. Feng, Y., Ueda, A. & Wu, C.-F. A modified minimal hemolymph-like solution, HL3.1, for physiological recordings at the neuromuscular junctions of normal and mutant Drosophila larvae. *J. Neurogenet.* **18**, 377–402 (2004).

31. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

32. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

33. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

34. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

35. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation.

36. Hwang, R. Y. *et al.* Nociceptive Neurons Protect Drosophila Larvae from Parasitoid Wasps. *Curr. Biol.* **17**, 2105–2116 (2007).

37. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).

38. Xiang, Y. *et al.* Light-avoidance-mediating photoreceptors tile the Drosophila larval body wall. *Nature* **468**, 921–926 (2010).

39. Honjo, K., Hwang, R. Y. & Tracey, W. D. Optogenetic manipulation of neural circuits and behavior in Drosophila larvae. *Nat. Protoc.* **7**, 1470–1478 (2012).

40. Turner, H. N. *et al.* The TRP Channels Pkd2, NompC, and Trpm Act in Cold-Sensing Neurons to Mediate Unique Aversive Behaviors to Noxious Cold in Drosophila. *Curr. Biol. CB* **26**, 3116–3128 (2016).

41. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

42. Terada, S.-I. *et al.* Neuronal processing of noxious thermal stimuli mediated by dendritic Ca(2+) influx in Drosophila somatosensory neurons. *eLife* **5**, (2016).

43. Yan, Z. *et al.* Drosophila NOMPC is a mechanotransduction channel subunit for gentle-touch sensation. *Nature* **493**, 221–225 (2013).

44. Hu, C. *et al.* Sensory integration and neuromodulatory feedback facilitate Drosophila mechanonociceptive behavior. *Nat. Neurosci.* **20**, 1085–1095 (2017).

45. Ohyama, T. *et al.* A multilevel multimodal circuit enhances action selection in Drosophila. *Nature* **520**, 633–639 (2015).

46. Zhang, W., Yan, Z., Jan, L. Y. & Jan, Y. N. Sound response mediated by the TRP channels NOMPC, NANCHUNG, and INACTIVE in chordotonal organs of Drosophila larvae. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13612–13617 (2013).

47. Rosenzweig, M. *et al.* The Drosophila ortholog of vertebrate TRPA1 regulates thermotaxis. *Genes Dev.* **19**, 419–424 (2005).

48. Lee, Y. *et al.* Pyrexia is a new thermal transient receptor potential channel endowing tolerance to high temperatures in Drosophila melanogaster. *Nat. Genet.* **37**, 305–310 (2005).

49. Jang, W. *et al.* The anoctamin family channel subdued mediates thermal nociception in Drosophila. *J. Biol. Chem.* **290**, 2521–2528 (2015).

50. Yoshino, J., Morikawa, R. K., Hasegawa, E. & Emoto, K. Neural Circuitry that Evokes Escape Behavior upon Activation of Nociceptive Sensory Neurons in Drosophila Larvae. *Curr. Biol. CB* **27**, 2499–2504.e3 (2017).

51. Burgos, A. *et al.* Nociceptive interneurons control modular motor pathways to promote escape behavior in Drosophila. *eLife* **7**, (2018).

52. Jovanic, T. *et al.* Competitive Disinhibition Mediates Behavioral Choice and Sequences in Drosophila. *Cell* **167**, 858–870.e19 (2016).

53. Trinh, K. & Storm, D. R. Detection of odorants through the main olfactory epithelium and vomeronasal organ of mice. *Nutr. Rev.* **62**, S189-192-241 (2004).

54. Hilliard, M. A., Bargmann, C. I. & Bazzicalupo, P. C. elegans responds to chemical repellents by integrating sensory inputs from the head and the tail. *Curr. Biol. CB* **12**, 730–734 (2002).

55. Fettiplace, R. Hair Cell Transduction, Tuning, and Synaptic Transmission in the Mammalian Cochlea. *Compr. Physiol.* **7**, 1197–1227 (2017).

56. Obholzer, N. *et al.* Vesicular glutamate transporter 3 is required for synaptic transmission in zebrafish hair cells. *J. Neurosci. Off. J. Soc. Neurosci.* **28**, 2110–2118 (2008).

57. Roper, S. D. & Chaudhari, N. Taste buds: cells, signals and synapses. *Nat. Rev. Neurosci.* **18**, 485–497 (2017).

58. Stebbings, L. A., Todman, M. G., Phelan, P., Bacon, J. P. & Davies, J. A. Two Drosophila innexins are expressed in overlapping domains and cooperate to form gap-junction channels. *Mol. Biol. Cell* **11**, 2459–2470 (2000).

59. Ohno, Y. & Otaki, J. M. Spontaneous long-range calcium waves in developing butterfly wings. *BMC Dev. Biol.* **15**, 17 (2015).

60. Moe, A. M., Golding, A. E. & Bement, W. M. Cell healing: Calcium, repair and regeneration. *Semin. Cell Dev. Biol.* **45**, 18–23 (2015).

61. Toyota, M. *et al.* Glutamate triggers long-distance, calcium-based plant defense signaling. *Science* **361**, 1112–1115 (2018).

6.7 Chapter Six Figures

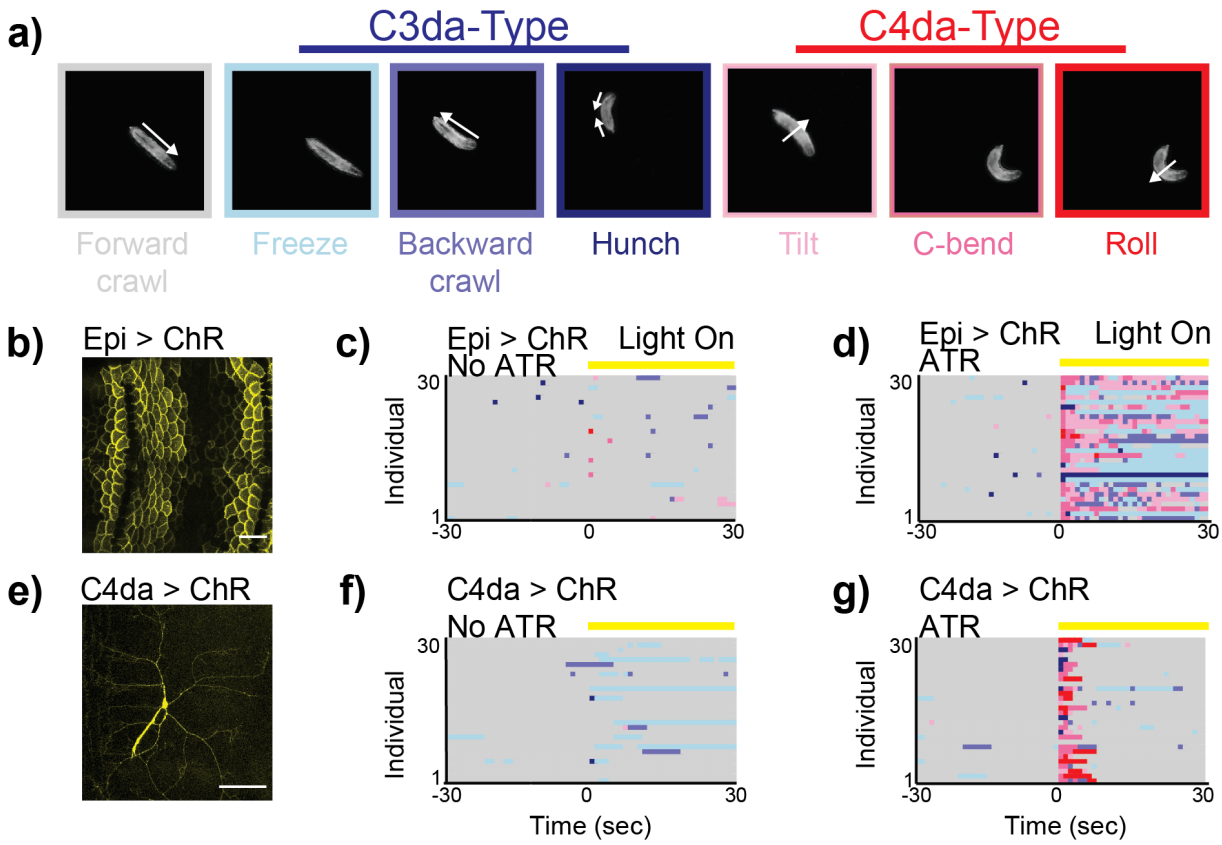

Figure 6.1 Epidermal stimulation induces nociceptive behaviors

(a) Still frames show different larval behaviors observed. (b) Epidermal (Epi) expression, driven by *A58-Gal4, tsh-Gal80, cha-Gal80*, of the fluorescently-tagged optogenetic channel *UAS-CsChrimson-mVenus* (ChR) in the larval body wall. (c, d) Ethogram of behaviors observed following red light (585 nm) activation of epidermal ChR in the absence (c) or presence (d) of the obligate cofactor all trans-retinal (ATR). (e) Nociceptor expression, driven by *ppk-Gal4*, of ChR in larval C4da somatosensory neurons. (f, g) Ethogram of behaviors observed following light activation of nociceptor ChR in the absence (f) or presence (g) of ATR. Scale bars, 100μm. Colors in ethograms correspond to behaviors as indicated in (a). Each row of each ethogram corresponds to an individual animal. N = 30 animals per panel. Times are relative to the start of the light stimulus.

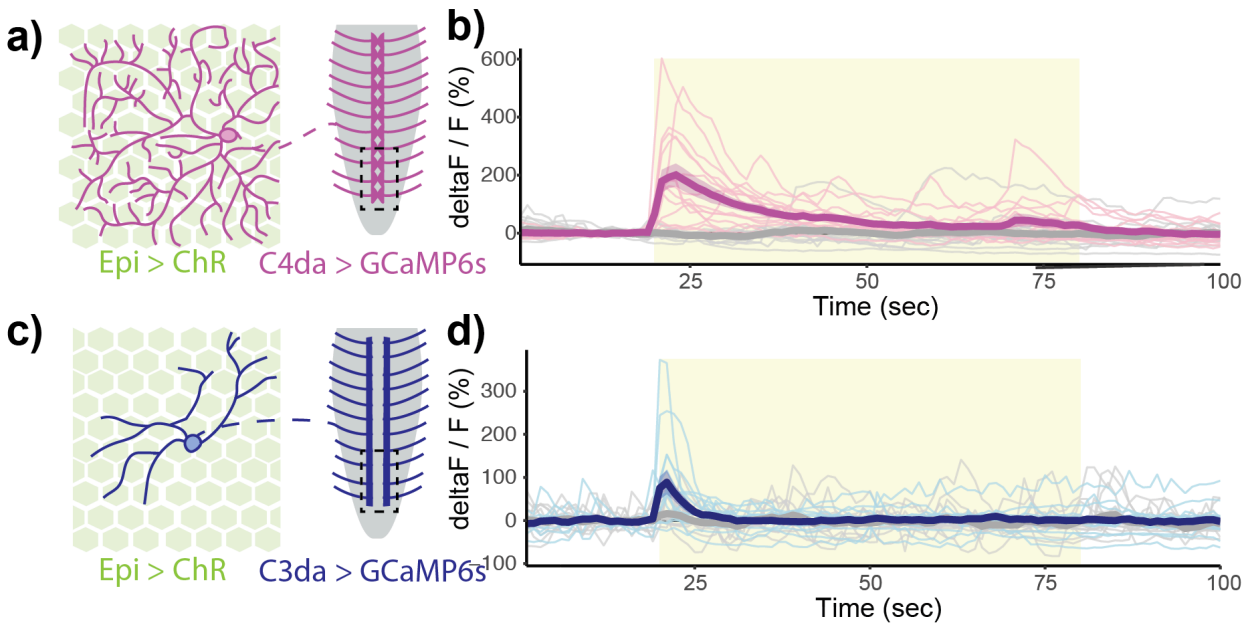

Figure 6.2 Epidermal stimulation induces activation of somatosensory neurons

(a) Schematic of C4da imaging setup. CsChrimson (ChR) is expressed in the epidermis and calcium imaging of GCaMP6s fluorescence is performed in the axon terminals of somatosensory neurons. Dotted line indicates sample imaging window. (b) Line plot showing the change in fluorescence of GCaMP6s in C4da neurons during epidermal stimulation. (c) Schematic of C3da imaging setup. (d) Line plot showing the change in fluorescence of GCaMP6s in C3da neurons during epidermal stimulation. Colored lines (rmagentas, blues), ATR plus. Grey lines, ATR minus. Bold lines, mean. Shading around mean, SE. Thin lines, traces from individual samples. N = 15 samples per condition.

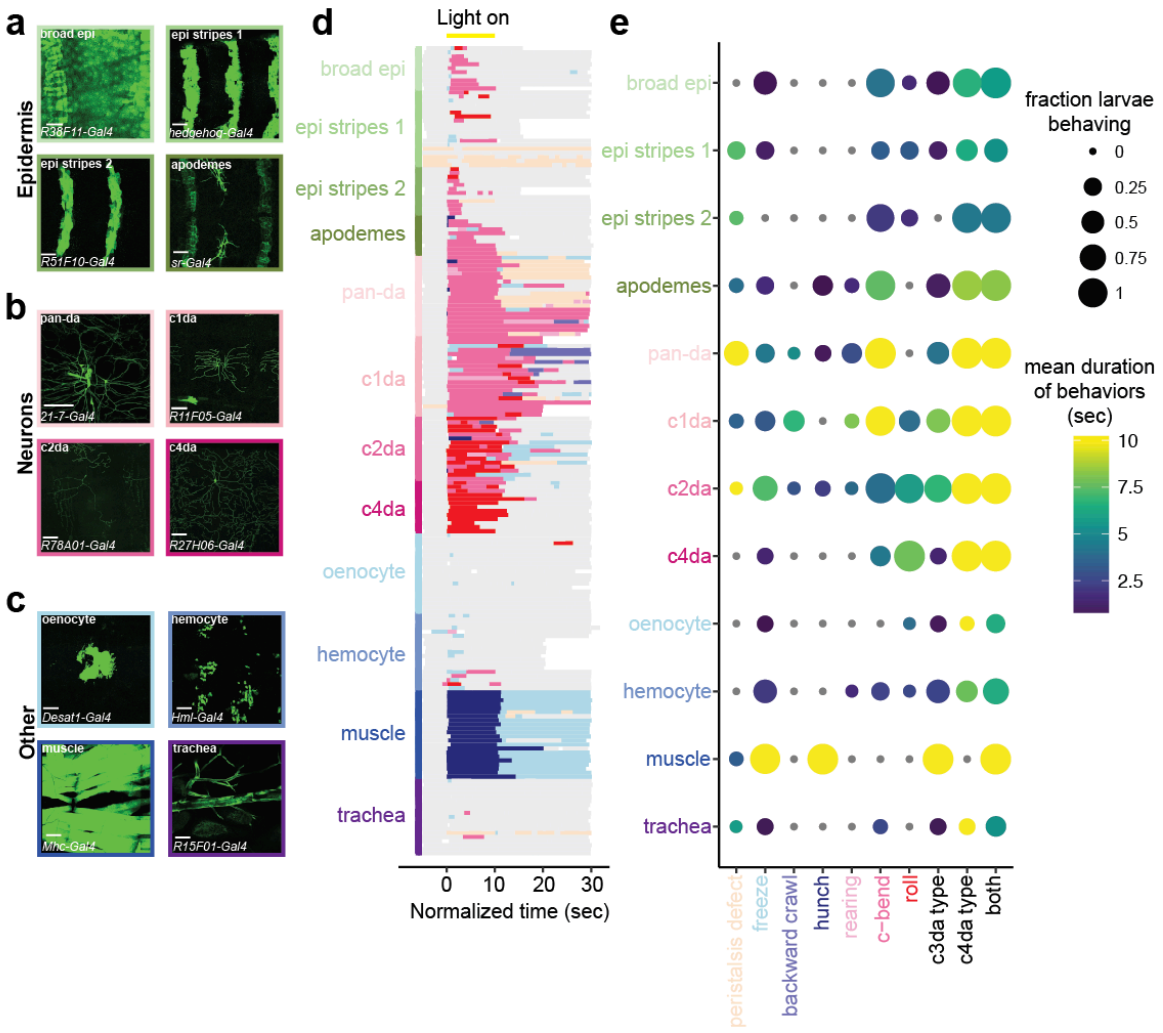

Figure 6.3 Induction of nociceptive behaviors following stimulation is unique to epidermal subtypes among peripheral non-neuronal cell types

Expression pattern, shown with cytosolic *2xeGFP*, of four epidermis-specific drivers (a), four neuron-specific drivers (b), and four other peripheral cell type drivers (c). Scale bars, 100μm. (d) Ethogram of behaviors observed during green light (488 nm) activation of *CsChrimson* in each of the indicated cell types. N ≥ 10 larvae per cell type. Times are relative to the start of the light stimulus. Behavior colors correspond to the behavior labels along the x-axis in panel (e). (e) Dot plot shows the fraction of larvae (dot size) exhibiting each behavior and mean duration (dot color) of each behavior among those exhibiting the behavior for each cell type. The last three columns summarize C3da-type behaviors (freeze, backwards crawl, hunch), C4da-type behaviors (rear, c-bend, roll), and all responsive behaviors.

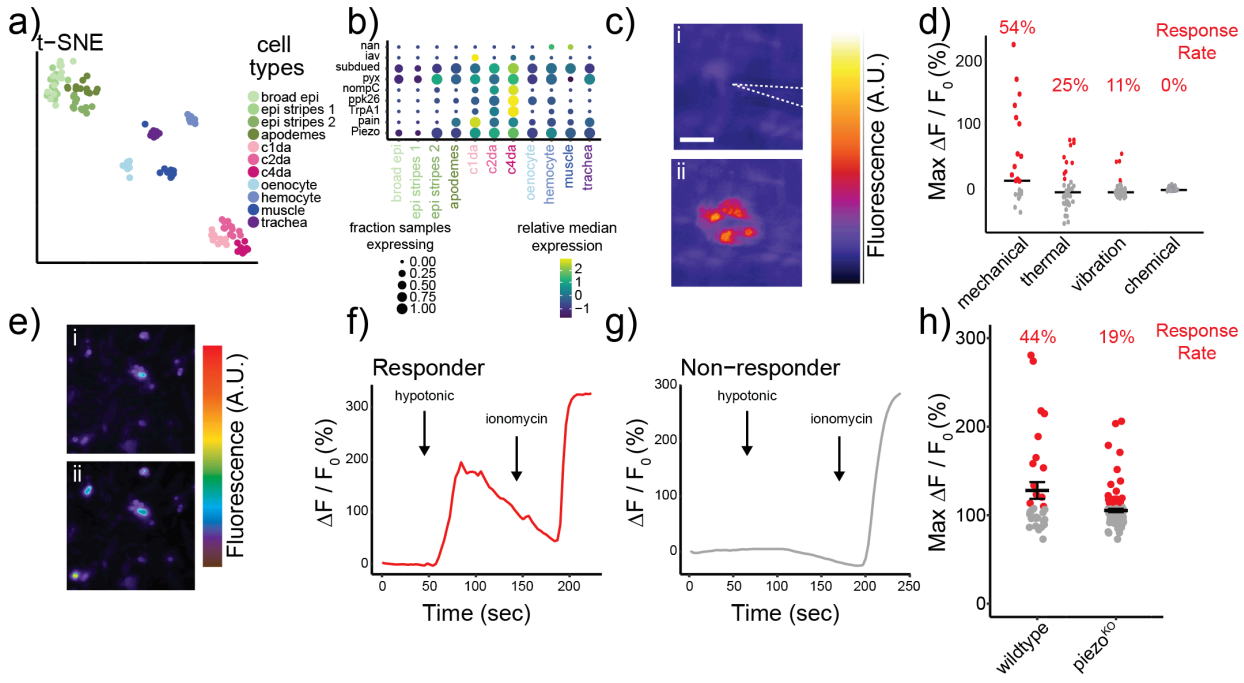

Figure 6.4 Epidermal cells are intrinsically mechanosensitive

(a) T-SNE visualization of all transcriptomes profiled for different cell types in the larval periphery. Each point represents a single sample of 10 cells. (b) Dot plot of the expression of known sensory receptors across profiled cell types. Mean log expression, shown on the color scale, is normalized by gene. (c) Expression pattern of GCaMP6s in the epidermis in a larval filet showing fluorescence intensity prior to (i) and 20 seconds following (ii) a mechanical stimulus. Dotted line represents the position of mechanical probe. (d) Dot plot showing the maximum change in fluorescence of GCaMP6s following the indicated stimuli. Each point represents a single epidermal cell and bars represent median change. Mechanical, 10μm displacement. Thermal, 43°C. Vibration, 500Hz. Chemical, 100mM AITC. Red indicates a responding cell with at least 10% increase over baseline. (e) Expression of GCaMP6s in dissociated, single epidermal cells showing fluorescence intensity prior to (i) and following (ii) a hypotonic stimulus. Scale bar, 100μm. Representative traces of change in fluorescence of single epidermal cells responding to (f) or not responding to (g) a hypotonic stimulus. (h) Dot plot showing the maximum change in expression of GCaMP6s following the hypotonic stimulus in wildtype or *piezo*^{KO} backgrounds. Each point represents a single epidermal cell and lines represent mean change +/- SEM. Red indicates a responding cell with at least 10% increase over baseline. N ≥ 32 cells per genotype. A.U., arbitrary units.

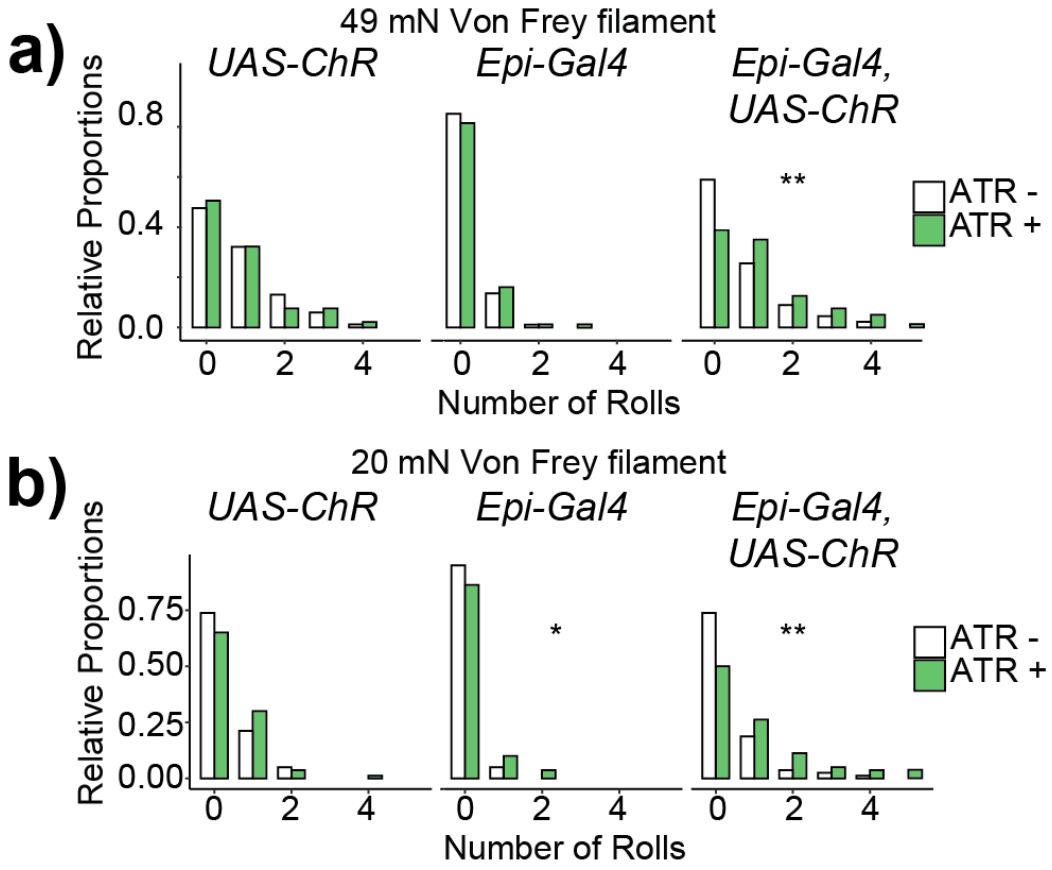

Figure 6.5 Epidermal activation state is behaviorally relevant

Histograms show the number of nociceptive rolls in response to a 49mN (a) or 20mN (b) von Frey filament across different genotypes, in the presence (green) or absence (white) of the CsChrimson co-factor ATR. Roll data were fit with a Poisson model to assess significance between ATR positive and ATR negative conditions for each genotype, followed by Benjamini-Hochberg correction. ChR, *CsChrimson*. Epi, epidermal. *, $p < 0.05$. **, $p < 0.01$.

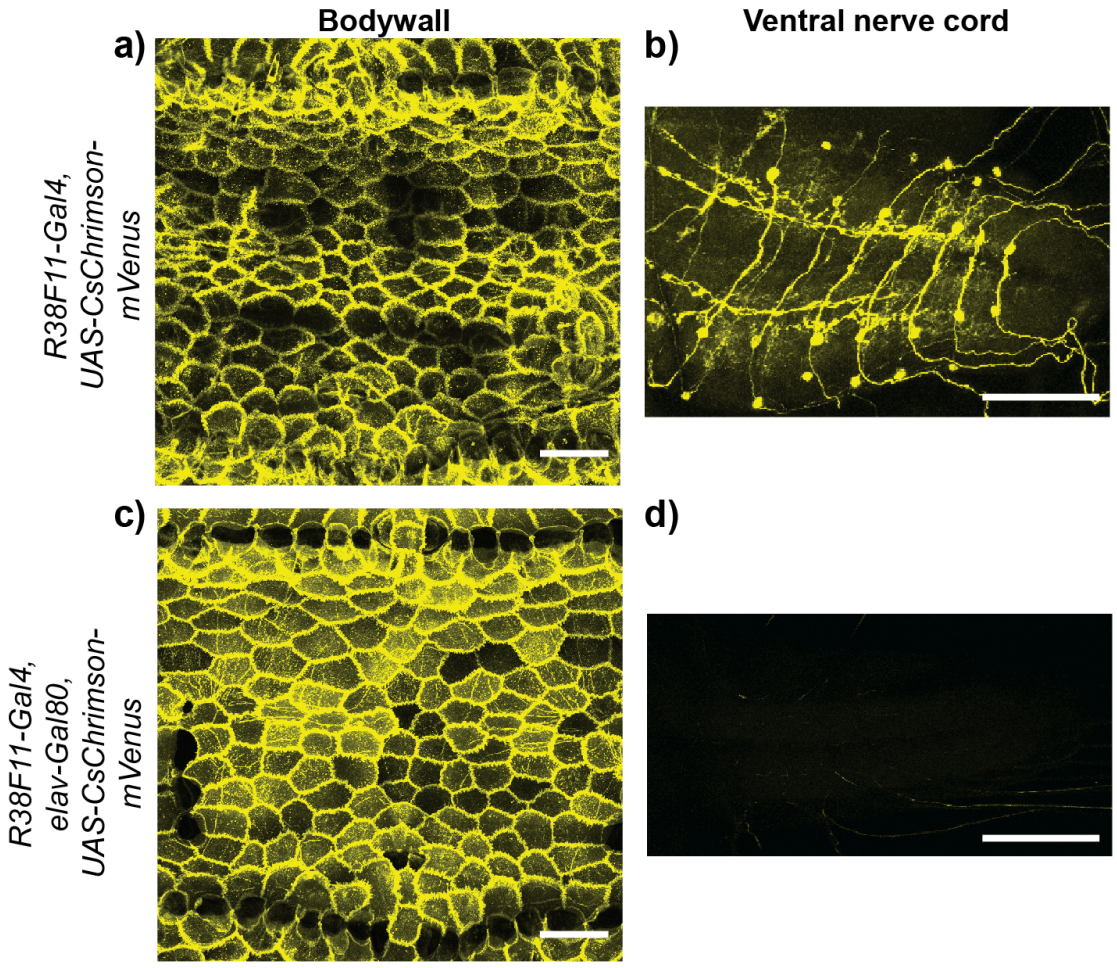

Figure 6.6 S1 Epidermal driver expression patterns

Maximum intensity projection images show the expression pattern of *R38F11-Gal4, UAS-CsChrimson-mVenus* in the body wall (a, c) or the ventral nerve cord (b, d) of larvae in the absence (a, b) or presence (c, d) of *elav-Gal80* to suppress neuronal expression. Scale bars, 100µm.

6.8 Chapter Six Tables

Table 6.1 Criteria for behavior scoring

Behavior	Description
Locomotion	Continuous movement with smooth, regular peristalsis from posterior to anterior at approximately the rate the larva exhibited prior to stimulus
Turning	Head turns at an angle relative to body, followed by movement of the rest of the body in the direction of the head. Turning terminates and turns into locomotion once the body is straight in the new direction. Turning requires peristalsis as or immediately after the animal is moving its head and changing direction (to differentiate from head sweeping, where the body is still)
Head sweep	Head moves in one direction and then the other at least once while the animal is not moving or exhibiting peristalsis.
Peristalsis defect	Clear defect in peristalsis. Typically determined relative to animal's locomotion prior to stimulus, taking that as baseline. Threshold is quite severe, such as slow or incomplete peristalsis, or peristalsis without locomotion, or peristalsis in opposite direction (without backwards movement).
Rolling	Larva completes a full 360-degree rotation of its body. Marked at initiation of the first full rotation, and stops at the end of the last full rotation before the animal's movement does not result in another complete roll.
C-bend	Bending into a c-shape, typically including bending of both head and tail in same direction. Also includes associated nocifensive behaviors such as writhing, whipping back and forth of head, tilting back and forth on the long axis without a complete 360-degree rotation (i.e. incomplete roll), and tail sweeping back and forth
Rearing	Upwards bending or arching of just the anterior part of the worm. Animal is typically stationary in this position, whereas if it is moving back and forth from this position or writhing about it would be characterized as c-bending.
Hunch	Contraction of body as larva remains completely still. Not a hunch if animal contracts on its way to a c-bend or roll in a continuous manner.
Backwards	Anterior-to-posterior peristalsis accompanied by movement in reverse.
Freeze	Stopping. Must be > 1 second.
Pause	Momentary halt, < 1 second.

Table 6.2 Solution compositions

Compound	HL3 Concentration (mM)	Hypotonic HL3 Concentration (mM)	HL3.1 Concentration (mM)	Calcium-free HL3.1 Concentration (mM)
NaCl	70	70	70	70
KCl	5	5	5	5
NaHCO3	10	10	10	10
HEPES	5	5	5	5
Trehalose	5	5	5	5
Sucrose	115	16.6	115	115
CaCl2	1	1	1.5	0
MgCl2	10	10	4	4
pH	7.2	7.2	7.1	7.1
Osmolarity (mOsm / L)	328	229.6	311.5	307