COMPUTATIONAL APPROACHES TO BIOLOGICAL DATA WITH APPLICATIONS IN IMAGE ANALYSIS, HUMAN VARIANT PRIORITIZATION, AND METAGENOMICS

by

Steven Flygare

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

August 2015

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The following faculty members served as the supervisory committee chair and members for the dissertation of______Steven Flygare__________________.

Dates at right indicate the members' approval of the dissertation.

_______Mark Yandell______________________, Chair __6/12/2015______
Date Approved

_______Lynn Jorde_________________________, Member ___6/12/2015 _____
Date Approved

_______Ellen Pritham_______________________, Member ___6/8/2015_______
Date Approved

_______Michael Shapiro _____________________, Member ___6/10/2015______
Date Approved

_______Christopher Gregg __________________, Member ___6/10/2015______
Date Approved

The dissertation has also been approved by______Lynn Jorde___________________ Chair of the Department of _________Human Genetics______________________

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Advances in technology have produced efficient and powerful scientific instruments for measuring biological phenomena. In particular, modern microscopes and next-generation sequencing machines produce data at such a rate that manual analysis is no longer practical or feasible for meaningful scientific inquiries. Thus, there is a great need for computational strategies to organize and analyze huge amounts of data produced by biological experiments. My work presents computational strategies and software solutions for application in image analysis, human variant prioritization, and metagenomics.

The information content of images can be leveraged to answer an extremely broad spectrum of questions ranging from inquiries about basic biological processes to highly specific, application-driven inquiries like the efficacy of a pharmaceutical drug. Modern microscopes can produce images at a rate at which rigorous manual analysis is impossible. I have created software pipelines that automate image analysis in two specific applications domains. In addition, I discuss general image analysis strategies that can be applied to a wide variety of problems.

There are tens of millions of known human genetic variants. Prioritizing human variants based on how likely they are to cause disease is of huge importance because of the potential impact on human health. Current variant prioritization methods are limited by their scope, efficiency, and accuracy. I present a variant prioritization method, the VAAST variant prioritizer, which is superior in its scope, efficiency, and accuracy to existing variant prioritization methods.

The rise of next-generation sequencing enables huge quantities of sequence to be generated in a short period of time. No field of study has been affected by rapid sequencing more than metagenomics. Metagenomics, the genomic analysis of a population of microorganisms, has important implications for pathogen detection because metagenomics enables the culture-free detection of microorganisms. I have created Taxonomer, a comprehensive metagenomics pipeline that enables the real-time analysis of read datasets derived from environmental samples.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Mark Yandell, for the outstanding mentor he has been and for the excellent research projects he had me work on during my graduate career. Without his mentoring and scientific advisement, my graduate work would not have been possible. I would also like to thank all the co-authors on my publications for their help and contributions to my work. Specific thanks goes to Robert Schlaberg, Keith Simmon, Brett Kennedy, Hao Hu, Barry Moore, and Zev Kronenberg for all the work we did together. Thanks to my lovely wife Whitney and our son James for their love and support during this process. Finally, a special thanks is needed for my mom Lori Flygare for her continuous encouragement throughout my life and education.

CHAPTER 1

INTRODUCTION

Computational approaches to large-scale biological data

Increasingly, experiments in the biological sciences are producing data at a scale that cannot be analyzed manually, even with a team of scientists, and the rate of data production is expected to only increase (Jiang & Liu, 2015; Seife, 2015). While large amounts of data present many opportunities for scientific discovery, this data deluge presents scientists with many challenges. The challenges associated with dealing with massive amounts of data are intrinsically computational, and have created a rising importance of effective computational techniques to store, organize, and analyze data. My research focus has been to develop computational techniques to analyze large datasets (datasets of sufficient size as to be impractical to analyze manually) of biological interest. In my dissertation, I detail specific computational approaches and applications in image analysis, human genetic variant prioritization, and metagenomics.

Image analysis

Image analysis is becoming increasingly important in the biosciences. Image data provides a wealth of phenotype information that can be used to understand biological mechanisms in a wide range of applications, including experiments to uncover gene function or to determine the impact of a pharmaceutical drug (Carpenter et al., 2006). Increasingly sophisticated imaging techniques and microscopes produce quality data in such quantities that would take a team of researchers months to manually process the results of a single experiment. Thus, the potential impact of image analysis automation is enormous.

Image data acquired from experiments present many different challenges to an automated analysis. These challenges include the deep complexity represented in images, image quality, cell boundaries that are not completely defined, asymmetrical illumination, small sample sizes, high dimensionality, and small effect sizes between experimental groups of interest. These challenges together with the amount of data that needs to be processed present a significant computational challenge.

Because of the focused nature of most experiments, there is no single analysis pipeline that will work to analyze the images and produce meaningful statistics for all experiments. Thus, it is necessary to understand both image analysis methods and the statistics used to process the resulting data in order to draw meaningful conclusions from images produced by biological experiments. There are, however, existing image analysis software that is both modular and designed to allow experimental scientists (not just computational experts) to analyze their data. Examples include CellProfiler and ImageJ (Carpenter et al., 2006; Collins, 2007). Although these software packages exist, it is my belief that a user must have at least a conceptual understanding of the methods employed in order to direct an analysis and draw meaningful conclusions from images. I opt to use the excellent open source image analysis libraries available for the Python programming language and construct custom image analysis pipelines. These open source libraries include ndimage in SciPy, Scikit-Image, Python Imaging Library (PIL), Mahotas, and OpenCV. These libraries include excellent implementations of most major image analysis algorithms and are typically designed to work on numpy arrays for speed. Chapter 2 describes an image analysis pipeline I constructed using the Python programming language to processes images of the flatworm *S. mediterranea*. Chapter 3 describes an application of image analysis to quantify muscle fiber cell size, for which I also constructed an analysis pipeline using Python to analyze the images and perform statistical analysis of the analyzed output.

Here I will give a high level description for conceptual understanding of a few fundamental image analysis procedures. These core image analysis procedures include image thresholding, erosion and dilation methods, and feature size and location quantification.

Thresholding

Image thresholding / binarization is the process of separating pixels into a foreground and background.

An example of image thresholding is shown in Figure 1.1. There are many thresholding methods to choose from, but they can be broken into two broad categories: global and local thresholding. Global thresholding methods choose a single pixel value with which to divide all the pixels of the image into foreground and background. Global thresholding can be effective with relatively simple images where the lighting is uniform. However, global thresholding is ill suited when there is asymmetric illumination in an image, like that of Figure 1.1 A. In these cases, a local thresholding method is usually better suited. Local thresholding methods choose different thresholding values to use at different locations in the image. Figures 1.1 B and 1.1 C are the results of different local thresholding methods. Clearly, the method of Figure 1.1 B is superior in this application to that of Figure 1.1 C. Local thresholding methods can be broken into two categories: Scale-dependent and scale-independent methods (Blayvas, Bruckstein, & Kimmel, 2006). Scale-dependent methods have a specified neighborhood size around each pixel that is used to calculate a local threshold. Fox example, we may consider a 20 x 20 box of pixels around every pixel to be its neighborhood and use the pixel information of the neighborhood to calculate a threshold value for the particular pixel. Scale-independent methods do not specify any particular neighborhood size around a pixel; instead, they typically combine pixel intensity measures for regions of many different sizes around the pixel. Scale-dependent methods can be very effective in solving problems when there is an expectation about the size of the objects of interest. Scale-dependent methods also have the advantage of being simpler to understand and implement.

Figure 1.2 A is an image taken by a BD Pathway Bioimager of the flatworm S. mediterranea. The purpose of the experiment that produced these images was to quantify the neoblasts in mutant animals produced by an RNAi screen and compare the neoblast counts to control animals. The neoblasts are stained prior to imaging so they become the brightest points of light in the image. I used a scale-dependent method to threshold these images because of

asymmetric illumination produced by the microscope with the some of the images. Figure 1.2 B shows the results of this thresholding method – you can see the neoblasts were easily separated from the image background using this thresholding technique.

In my experience, there is no single thresholding method that is going to work for all images. I recommend testing a few methods, including both global and local, scale-dependent and scale-independent, on a few of your images and selecting the method that works best for your particular data.

Erosion and dilation

Once an image is thresholded adequately, it becomes possible to count and quantify features in the image. Often times, the features of interest in an image are not completely separate in the image after thresholding and need to be separated before quantifying their size. For example, Figure 1.3 is an image taken by a confocal microscope of the cross section of a mouse Tibialis anterior muscle. Our purpose in analyzing this image is to quantify the size of the muscle fibers, which in Figure 1.3 are outlined by the red channel.

Applying a thresholding procedure to Figure 1.3 results in Figure 1.4 A. Thresholding the image does not provide enough separation between the muscle fibers to quantify their size because many of the fibers are still touching. Erosion is a process that shrinks features in the image and thereby enables the separation of the features. Applying one erosion step to Figure 1.4 A results in Figure 1.4 B and applying two erosion steps to Figure 1.4 A results in Figure 1.4 C. The fibers in both Figure 1.4 B and 1.4 C look separate enough to do quantification. In general, when using erosion to isolate features as we have done here, it is desirable to do the minimum amount of erosion necessary to isolate the features. By using the least amount of erosion, we are able to use the maximum amount of image data. If the experiment were to include comparing muscle fiber size between groups of animals, it would be critically important to use the same erosion steps when doing the image analysis since erosion systematically changes the measurable size of the muscle fibers.

Feature size and location quantification

After an image has been thresholded and appropriate erosion steps have been taken to isolate the features of interest, it is possible to quantify the size and location in the image of each of the features. In the case of the muscle fiber image shown in Figure 1.3, the objective is to quantify the size of each of the muscle fibers (outlined by red). Once the image looks like Figure 1.4 B or 1.4 C, quantification can take place. Here I will give a short description of a common method used to quantify the size of isolated features. This method begins by selecting a pixel that is above the threshold (white pixels in Figures 1.4 A, B, C) and then looks at all of its neighbors – every pixel has 8 neighbors. For every neighbor that is a foreground pixel, this process is repeated for each neighbor until no more neighboring foreground pixels are found. These pixels are saved as a single feature and this process is repeated until no more foreground pixels are left in the image. We now have a collection of pixels grouped by feature. At this point, we know the size of each feature in pixels. In addition, by taking the average of the x and y coordinates of each pixel of a feature we find its center of mass, which is often a location quantity of interest. It is important to note that the center of mass thus calculated can be different from the visual center of a feature. An example is of a banana shaped feature – its center of mass would lie outside the feature.

Human variant prioritization

Over the past decade, sequencing costs have dropped precipitously. The super-exponential drop in sequencing costs has led to a massive increase in sequencing-related research and applications (Katsonis et al., 2014). This ever-increasing wealth of sequence data has resulted in an explosion of known human variants. For example, the NCBI's dbSNP database contains well over 100 million human variants. This available panoply of human variation presents significant challenges to interpretation, and of particular importance is how to rank human variants according to their risk for causing or contributing to disease.

SIFT and PolyPhen were among the first recognized methods to prioritize human variants and are still viewed as a standard for variant prioritization (Ng & Henikoff, 2003; Ramensky, Bork,

& Sunyaev, 2002). SIFT uses information about amino acid conservation and the biochemical properties of the amino acids to assign a score to the observed nonsynonymous substitution. Like SIFT, PolyPhen is informed using amino acid conservation information, but in addition, PolyPhen also incorporates information about protein structure to score nonsynonymous substitutions. SIFT and PolyPhen still compare favorably to many methods that have since been developed to prioritize nonsynonymous amino acid changes (Dong et al., 2014).

Both SIFT and PolyPhen prioritize only nonsynonymous variants. In real applications, this limitation is extremely problematic since the vast majority of known human genetic variation is noncoding, and there are many known disease-causing variants in humans that fall outside the category of nonsynonymous protein coding change (Ritchie, Dunham, Zeggini, & Flicek, 2014). Prioritization of noncoding variants is a much more difficult problem than prioritization of nonsynonymous variants because there is comparably much less information available in noncoding regions. However, projects like ENCODE are attempting to functionally annotate noncoding regions by systematically assaying all functional genomic elements (Dunham et al., 2012).

Methods are needed that can accurately prioritize both coding and noncoding human genetic variation. Kircher et al. developed CADD, a machine learning approach to human variant prioritization that can score all SNVs and small indels in the human genome and is more effective than existing methods for variant prioritization (Kircher et al., 2014). CADD works by comparing incidence of simulated variants to fixed derived alleles in the human lineage. This clever comparison allows them to quantify the depletion of fixed derived alleles in the human lineage for all locations in the genome. The main idea is that genomic locations that have a relative depletion for fixed variation in the human lineage are more likely to have a functional consequence. However, CADD cannot score larger indels or other structural variation.

I have developed a variant prioritization method based on the VAAST likelihood, and in contrast to other available methods, it is able to prioritize all annotated variation across the human genome (Hu et al., 2013; Yandell et al., 2011). This method is called the VAAST Variant Prioritizer (VVP). The core concept behind VVP is to calculate a score for a variant that indicates how potentially damaging it is. This score is then compared to scores of known healthy human variants and its percentile rank is calculated. A high percentile rank (> 99) indicates that the variant looks more damaging than the majority of known healthy human variation. Implicit to this method is the problem of choosing how to organize healthy human variants into 'lookup' bins against which variants can be compared. Empirically, I have found that creating separate lookups for a set of user-specified annotated genomic features (usually genes) and then further segmenting the lookups into coding and noncoding categories produces an effective and efficient way to prioritize human variants. Details of VVP and its performance characteristics, including comparisons to CADD, are given in Chapter 4.

Metagenomics

Metagenomics is the genomic analysis of a population of microorganisms (Handelsman, 2004). Metagenomic analysis involves extracting DNA or RNA from an environmental sample, sequencing it, and using the sequence reads to identify organisms present in the sample.

The majority of microorganisms cannot be grown in a laboratory, but through metagenomic analysis, these microorganisms can be observed and studied since culturing is not required. For this reason, metagenomics holds incredible promise in terms of the possible questions it opens to investigation (Brady & Salzberg, 2009).

With falling sequencing costs, metagenomics projects have produced huge amounts of sequence data (Wood & Salzberg, 2014). The goal of a metagenomic analysis is to classify every read with as much taxonomic precision as possible. Blast is an extremely effective tool for comparing a query sequence to a database in order to produce a taxonomic classification, and is the standard of taxonomic classification accuracy. As such, the blast suite is the traditional choice for metagenomic analysis, but as sequence datasets have grown, blast is not fast enough to produce meaningful results in a reasonable amount of time (Wood & Salzberg, 2014).

Acquiring metagenomics results rapidly from an environmental sample has important consequences that because of the potential for real-time pathogen identification in response to disease outbreak and infections (Lipkin, 2013). Because metagenomics is hypothesis neutral, novel pathogens that contribute to disease can be identified, unlike the specific assays that are current medical practice for pathogen detection.

I have developed Taxonomer, a software pipeline for comprehensive metagenomic anlaysis. Taxonomer employs k-mer based methods to enable taxonomic classification based on rapid nucleotide and protein searches with a novel statistical approach that improves its accuracy over existing k-mer based methods while maintaining computational efficiency. Taxonomer also enables host transcription profiling. Full details and benchmarking of Taxonomer are given in Chapter 5.

K-mer based metagenomics

The need for metagenomic methods that are rapid enough to analyze the huge amount of sequence data has led to a proliferation of k-mer based methods. A k-mer is a k length substring of DNA sequence. For instance, the 3-mers of AAGGCGTC would be AAG, AGG, GGC, GCG, CGT, and GTC. Instead of using an alignment method that matches a seed (a k-mer) and then extends the alignment, k-mer based methods simply check for the presence or absence of a k-mer. This is a far more simple calculation than alignment seeding and extension; for this reason, k-mer-based methods can be hundreds or thousands of times faster than alignment based methods (Buchfink, Xie, & Huson, 2015; Patro, Mount, & Kingsford, 2014; Wood & Salzberg, 2014). Although the calculations in k-mer-based methods are simpler, the accuracy of read assignment from k-mer-based methods can be equivalent to that of the more computationally expensive alignment extension based approaches, even with sequencing errors (Buchfink et al., 2015; Edwards et al., 2012; Patro et al., 2014; Wood & Salzberg, 2014). In metagenomics, where rapid and accurate taxonomic assignment is more important than the information of a complete alignment, k-mer-based methods are the practical choice.

Database design

To unlock the speed of k-mer-based methods, careful database design and implementation choices are required. Here I will give an overview of the construction of a k-mer database for rapid queries, as well as a search strategy for k-mers. In order to create the database, all the k-mers in the reference sequences need to be identified. Effective software tools exist that will identify all the k-mers and their counts in a set of reference sequences, e.g., Jellyfish, Kanalyze, and KMC 2 (Audano & Vannberg, 2014; Deorowicz, Kokot, Grabowski, & Debudaj-Grabysz, 2015; Marçais & Kingsford, 2011). These k-mer counting tools all produce similar output tables of the k-mers and their counts; these tables can then be organized to allow for rapid k-mer queries. One possible organization of these tables for rapid queries depends on the concept of a k-mer minimizer (Figure 1.5) (Roberts, Hayes, Hunt, Mount, & Yorke, 2004). K-mers are organized into blocks based on a shared minimizer, and within the block, the k-mers are sorted in lexicographical order (Figure 1.6). An important observation is that overlapping k-mers often share the same minimizer (Wood & Salzberg, 2014). Since k-mers are organized into blocks by the minimizer they share, overlapping k-mers can first be searched in the minimizer block from the preceding k-mer and only calculate the minimizer if the k-mer is not found. Within a k-mer block, a binary search is used since the k-mers are in lexicographical order. This minimizer indexed query scheme produces astounding speeds even with extremely large datasets (Wood & Salzberg, 2014).

Another important implementation consideration to maximize speed is to represent k-mers as unsigned 64 bit integers; this can be achieved by using 2 bits to represent each of the 4 DNA base pairs. This numerical representation limits the length of k-mers to 31 bp in length, but is critical for good performance on large datasets. Implementation details of numerical k-mer representation are given in the papers describing Jellyfish, Kanalyze, and the source code of Kraken (Audano & Vannberg, 2014; Marçais & Kingsford, 2011; Wood & Salzberg, 2014).

References

Audano, P., & Vannberg, F. (2014). KAnalyze: A fast versatile pipelined K-mer toolkit. *Bioinformatics*, *30*(14), 2070–2072. doi:10.1093/bioinformatics/btu152

Blayvas, I., Bruckstein, A., & Kimmel, R. (2006). Efficient computation of adaptive threshold surfaces for image binarization. *Pattern Recognition*, *39*(1), 89–101. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.4468

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–6. doi:10.1038/nmeth.1358

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Under Review*, *12*(1). doi:10.1038/nmeth.3176

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., … Sabatini, D. M. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, *7*(10), R100. doi:10.1186/gb-2006-7-10-r100

Collins, T. J. (2007). ImageJ for microscopy. *BioTechniques*, *43*(1 Suppl), 25–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17936939

Deorowicz, S., Kokot, M., Grabowski, S., & Debudaj-Grabysz, a. (2015). KMC 2: Fast and resource-frugal k-mer counting. *Bioinformatics*, (January), 1–8. doi:10.1093/bioinformatics/btv022

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2014). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125–2137. doi:10.1093/hmg/ddu733

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., … Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Edwards, R. a, Olson, R., Disz, T., Pusch, G. D., Vonstein, V., Stevens, R., & Overbeek, R. (2012). Real time metagenomics : Using k -mers to annotate meta- genomes. *Bioinformatics (Oxford, England)*, 5–6.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. doi:10.1128/MBR.68.4.669

Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., & Yandell, M. (2013). VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, *37*(6), 622–34. doi:10.1002/gepi.21743

Jiang, P., & Liu, X. S. (2015). Big data mining yields novel insights on cancer. *Nature Genetics*, *47*(2), 103–104.

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations : Biological impact and theoretical interpretation. *Protein Science*, *23*, 1650–1666. doi:10.1002/pro.2552

Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–5. doi:10.1038/ng.2892

Lipkin, W. I. (2013). The changing face of pathogen discovery and surveillance. *Nature Reviews. Microbiology*, *11*(2), 133–41. doi:10.1038/nrmicro2949

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, *27*(6), 764–70. doi:10.1093/bioinformatics/btr011

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, *31*(13), 3812–3814. doi:10.1093/nar/gkg509

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–4. doi:10.1038/nbt.2862

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, *30*(17), 3894–3900. doi:10.1093/nar/gkf493

Ritchie, G. R. S., Dunham, I., Zeggini, E., & Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nature Methods*, *11*(3), 294–6. doi:10.1038/nmeth.2832

Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M., & Yorke, J. a. (2004). Reducing storage requirements for biological sequence comparison. *Bioinformatics*, *20*(18), 3363–3369. doi:10.1093/bioinformatics/bth408

Seife, C. (2015). The revolution is digitized. *Nature*, *518*, 480–481.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. doi:10.1186/gb-2014-15-3-r46

Yandell, M., Huff, C. D., Hu, H., Singleton, M., Moore, B., Xing, J., … Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, (21), 1529–1542. doi:10.1101/gr.123158.111

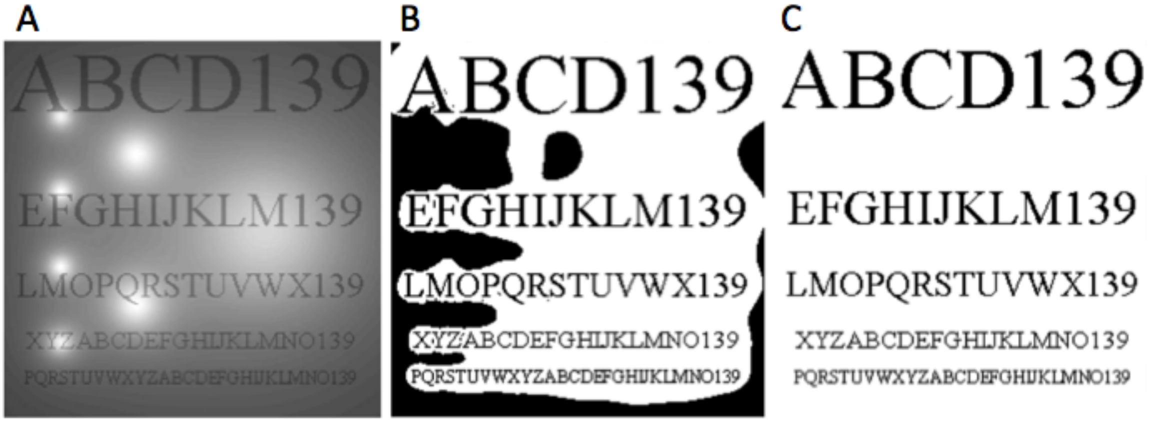


Figure 1.1: Original Image (A), note the assymetric illumination. The thresholding problem presented in (A) is to separate the letters from the rest of the image. Results of thresholding or binarization procedures (B,C). Different procedures yield better or worse results depending on the image, which is why its necessary to sample several procedures before choosing one for an analysis. Images taken and modified from (Blayvas et al., 2006).

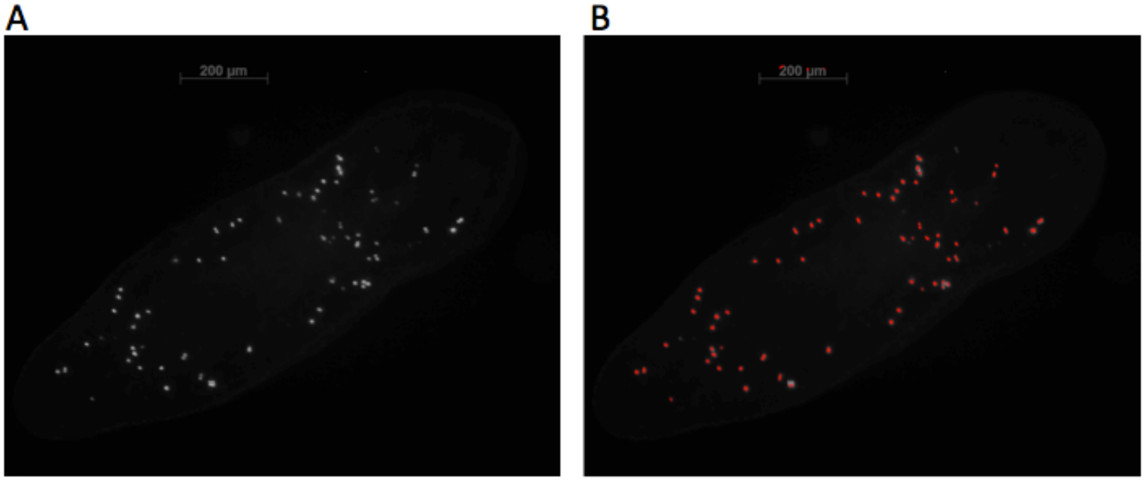


Figure 1.2: Image of S. meditteranea with stained neoblasts taken from a BD Pathway Bioimager (A). A scale-dependent thresholding method was able to effectively separate the neoblasts (shown in red) from the rest of the background (B).

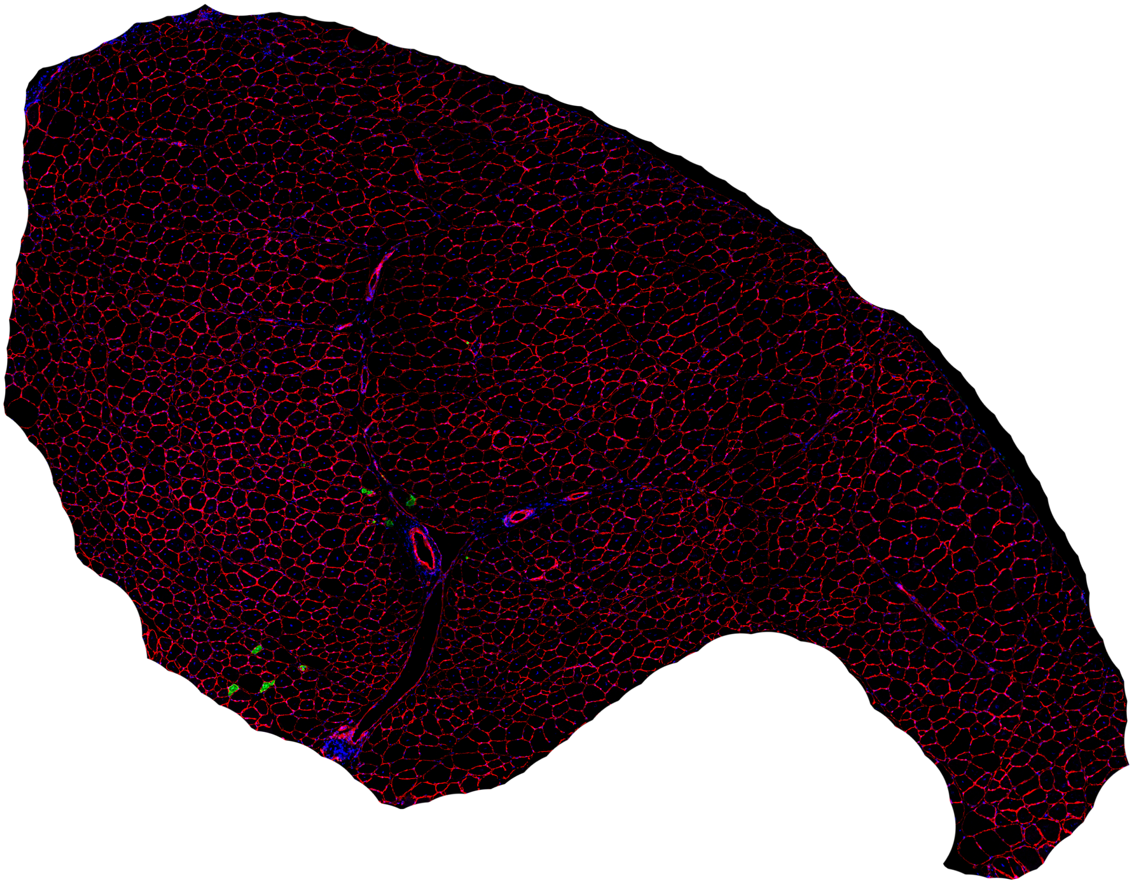


Figure 1.3: Stained cross section of the tibialis anterior muscle of a mouse. The red channel outlines the borders of the muscle fibers.

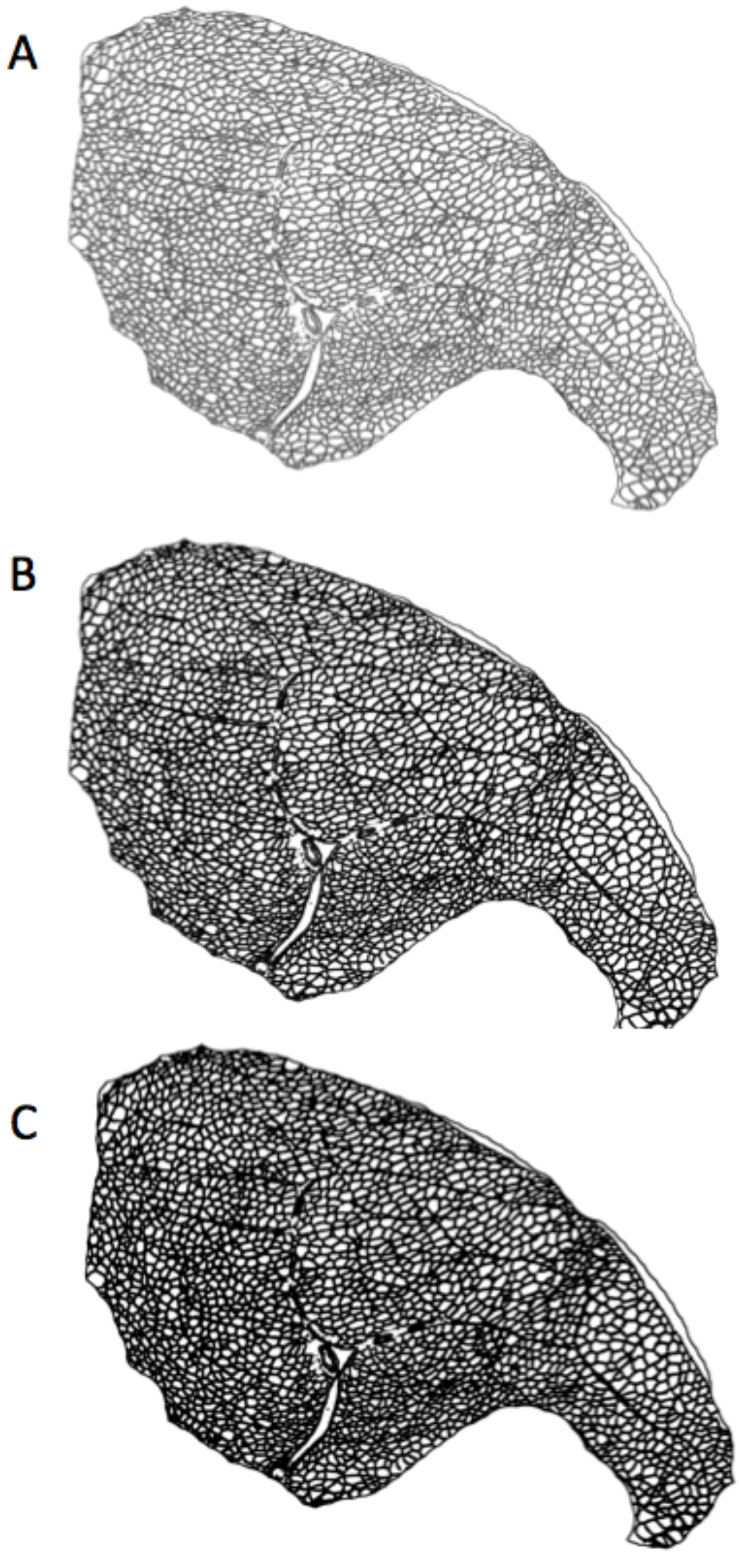


Figure 1.4: Impact of erosion on isolating muscle fibers. Thresholded image, no erosion (A). Thresholded image with one or two erosion steps (B,C).

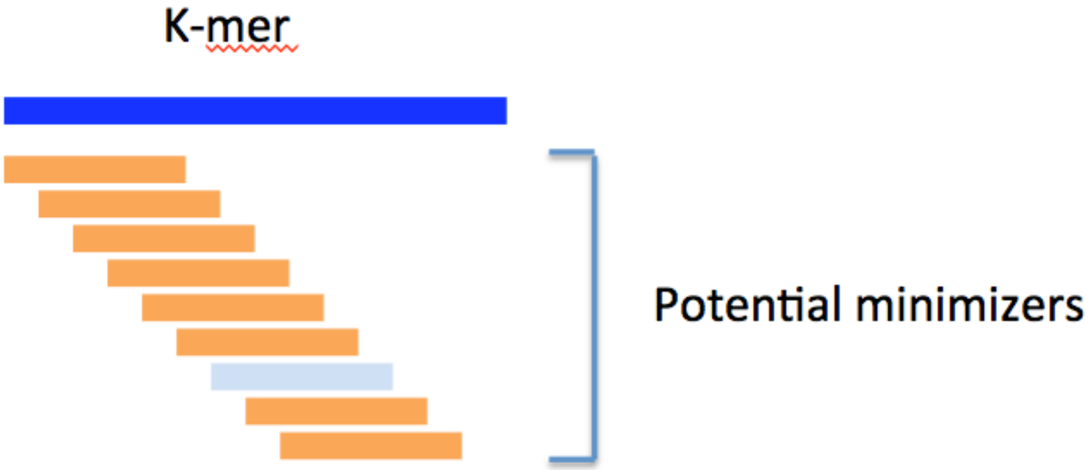


Figure 1.5: K-mer minimizer. To find the minimizer of a k-mer (shown in blue), all k-mers of a specified size smaller (shown in orange) than the original k-mer are generated from the k-mer in question. The k-mer minimizer (shown in light blue) is the potential minimizer that is the lexicographically smallest.

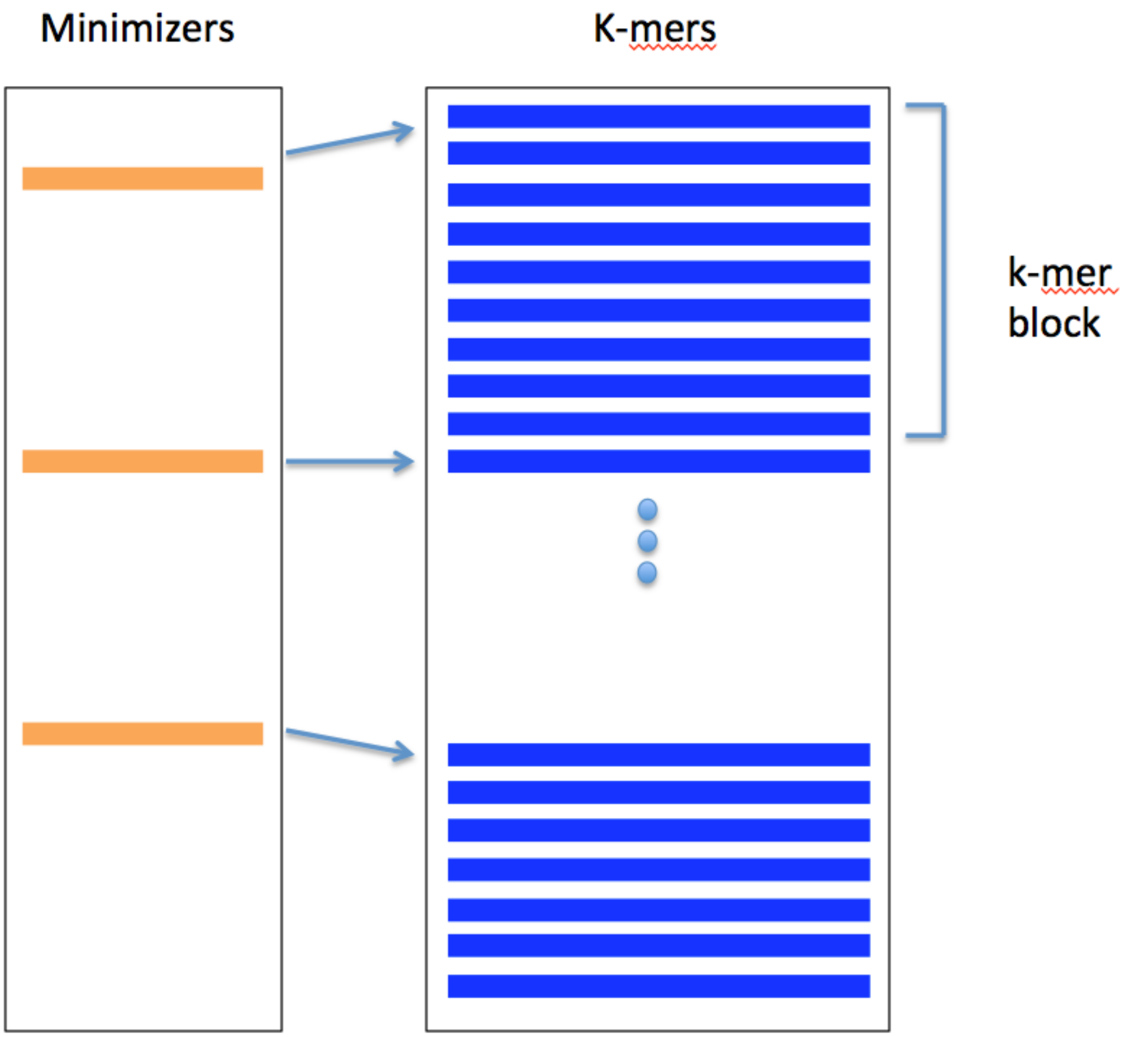


Figure 1.6: K-mer database organization by minimizer. K-mers (shown in blue) are organized into blocks based on shared minimizers. Minimizers (shown in orange) point to the beginning of k-mer blocks that are sorted in lexicographical order.

CHAPTER 2

IMAGEPLANE: AN AUTOMATED IMAGE ANALYSIS PIPELINE FOR HIGH-THROUGHPUT SCREEN USING THE PLANARIAN SCHMIDTEA MEDITERRANEA

Published as: Flygare, S., Campbell, M., Ross, R. M., Moore, B., & Yandell, M. (2013). ImagePlane: An Automated Image Analysis Pipeline for High-Throughput Screens Using the Planarian Schmidtea mediterranea. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *20*(8), 583–92. doi:10.1089/cmb.2013.0025

Contributions: I wrote the ImagePlane software, helped produce the images used in the analyses, was heavily involved in producing all the results, and wrote the methods.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 8, 2013
© Mary Ann Liebert, Inc.
Pp. 583–592
DOI: 10.1089/cmb.2013.0025

ImagePlane: An Automated Image Analysis Pipeline for High-Throughput Screens Using the Planarian *Schmidtea mediterranea*

STEVEN FLYGARE,* MICHAEL CAMPBELL,* ROBERT MARS ROSS,*
BARRY MOORE, and MARK YANDELL

ABSTRACT

ImagePlane is a modular pipeline for automated, high-throughput image analysis and information extraction. Designed to support planarian research, ImagePlane offers a self-parameterizing adaptive thresholding algorithm; an algorithm that can automatically segment animals into anterior–posterior/left–right quadrants for automated identification of region-specific differences in gene and protein expression; and a novel algorithm for quantification of morphology of animals, independent of their orientations and sizes. ImagePlane also provides methods for automatic report generation, and its outputs can be easily imported into third-party tools such as R and Excel. Here we demonstrate the pipeline's utility for identification of genes involved in stem cell proliferation in the planarian *Schmidtea mediterranea*. Although designed to support planarian studies, ImagePlane will prove useful for cell-based studies as well.

Key words: biology, functional genomics, genomics.

1. INTRODUCTION

The commercial availability of automated microscopes equipped with robotic sample-handling capabilities is making possible complex image-based assays that employ functional genomics techniques such as RNAi to investigate gene function at the genome scale (Kamath and Ahringer, 2003; Paddison and Hannon, 2003; Boutros et al., 2004; Kuttenkeuler and Boutros, 2004). These screens typically generate many thousands of images that must be processed and analyzed. Current paradigms of image processing and analysis generally involve graphical user interfaces (GUIs) to prepackaged collections of image-processing algorithms. Users typically employ these packages to process images one at a time or in batch mode using pull-down menus and check buttons. Although these software packages are useful, they also present researchers with practical difficulties when modification and customization are required. Obtaining the legal permissions and corporate support for customization, for example, is often a troublesome task. These have

Eccles Institute of Human Genetics, University of Utah, and School of Medicine, Salt Lake City, Utah.
*These three authors contributed equally to this work.

driven the development of publicly available image-processing libraries such as ImageJ (Collins, 2007; Papadopulos et al., 2007).

Although GUI-based packages are extremely useful, using them to process thousands of images even in batch mode can be time-consuming and exhausting. The large numbers of images generated by high-throughput image-based screens thus necessitate more automated approaches that minimize the need for GUI-mediated user interactions. Indeed, the ultimate goal of such automation (somewhat paradoxically) is the creation of image analysis pipelines that can rapidly extract information from large numbers of images without anyone ever actually looking at the images. In many respects, the challenges here resemble those previously encountered in the domain of genome annotation. Early genome annotation efforts were human-driven, with teams of investigators manually inspecting the details of aligned expressed sequence tags (ESTs) and proteins to a sequenced genome in order to deduce the intron–exon structures of novel genes (Oliver et al., 1992; Fleischmann et al., 1995). For reasons of economy and scale, the genomics field has gradually moved away from manual approaches, and today most genomes are annotated in an automated fashion (Curwen et al., 2004; Liang et al., 2009; Holt and Yandell, 2011). Today's image-based screens offer very similar challenges, and similar solutions are needed. A key point to appreciate in this regard is the distinction between solutions to basic problems in image processing—such as segmentation, registration, and thresholding—and the issues surrounding practical approaches to automated high-throughput image analysis. Like today's automated genome annotation pipelines (Curwen et al., 2004; Holt and Yandell, 2011), the challenge here is not so much to develop new techniques and algorithms, but rather to integrate existing tools and approaches into efficient, reliable, and accurate pipelines for automated information extraction and analyses of large collections of images.

Some of the most exciting opportunities for high-throughput image analyses involve screens of differentiating cells and embryos that employ RNAi and siRNA techniques to systematically perturb gene function at the genome scale. One problem here is the three-dimensionality of developing plant and animal embryos, which significantly complicates automated analyses. The need to register and segment images of un-orientated, morphologically complex embryos is a great challenge, one being addressed by many researchers today (Eliceiri et al., 2012). Although algorithmic breakthroughs are something to look forward to, there are other alternatives. One is to restrict the dimensionality of the problem by choosing less irregularly shaped organisms and tissues. In this regard, the planarian *Schmidtea mediterranea* is an obvious choice. Long renowned for its ability to regenerate, recent work has also shown that this planarian is an excellent model for stem cell biology. Equally important, planarians are literally flatworms. This fact greatly simplifies automated analyses. As our results demonstrate, *S. mediterranea* can be treated as two-dimensional for many image analysis applications; this has allowed us to largely circumvent the complexities associated with analyses of embryos and tissues having complex three-dimensional morphologies.

With these considerations in mind, we have developed an automated image analysis pipeline for planarian research called ImagePlane. This pipeline provides a self-configuring means to automatically threshold images, and to automatically identify and count stained cells. ImagePlane can also automatically segment images of planarians into anterior–posterior (A-P)/left–right (L-R) quadrants, a prerequisite for automated identification of region-specific differences in gene and protein expression. ImagePlane also provides a novel algorithm that allows rough, but rapid, quantification of morphological phenotypes independent of differences in animal orientation and size. This is an important step forward for planarian researchers, as animals can, and usually do, vary in size from individual to individual and between experimental batches. ImagePlane also provides practical methods for automatic report generation, and its outputs can be easily imported into third-party tools such as R and Excel. Here we demonstrate ImagePlane's utility using an image-based RNAi screen to identify genes involved in stem cell proliferation in the planarian *S. mediterranea*. Although designed primarily to support planarian studies, ImagePlane should prove useful for any high-throughput image-based investigation of approximately two-dimensional biological samples, including cell-based studies, and sectioned histological samples.

2. METHODS

2.1. Basic screen

We used a high-affinity antibody for phosphorylated histone 3 (H3P) (Millipore, Billerica, MA) to identify mitotic cells in *S. mediterranea.* Previous work has shown that this antibody provides effective means to identify neoblastic stem cells that are maintained throughout adult life (Hendzel et al., 1997; Newmark and Sánchez Alvarado, 2000; Reddien et al., 2005a). The screen proceeds as follows. First, animals are fed *E. coli* transformed with a plasmid designed to produce dsRNA of a chosen gene. Whole animals are fixed and stained with H3P antibody. *cdc23* was used as a positive RNAi control; this gene causes a twofold increase in the numbers of H3P-positive nuclei upon RNAi feeding (Reddien et al., 2005a). The *Caenorhabditis elegans unc-22* gene was used as a negative (placebo) control (Moerman et al., 1986; Yandell et al., 1994; Reddien et al., 2005a). Up to this point, this screen is identical to the one used by Reddien et al. (2005a). Next, animals were imaged using a BD Pathway Bioimager with a 10× objective. Animals were arrayed on 96-well plates with 44 RNAi-fed animals for three different genes, including 18 positive control (*cdc23*) and 15 negative control (*unc-22*) animals.

2.2. Determining animal outline and size

First, a simple algorithm is employed that iteratively computes an average weighted-by-pixel intensity to find the valley between the signal and background for each image. This algorithm works by starting with a guess, and then iteratively changing this guess until the number of background pixels weighted by intensity is equal to the foreground pixels weighted by intensity.

Next, each image is binarized by setting all the pixels below this value to 0 (minimum intensity) and above this value to 255 (maximum intensity). Ideally, every pixel with intensity above this threshold value is part of an animal. However, it is possible to have an image that has other objects (dust, etc.) outside the worm that also pass this particular thresholding filter. If objects exterior to the worm are not excluded, then the size of the worm will be miscalculated. Thus, objects exterior to the worm need to be differentiated. This is done using a recursive algorithm to determine which of the objects that passed the threshold is the largest. The following seven steps give a conceptual overview of how this is done. (1) Consider every pixel in the image as unvisited. (2) Move through the image-row by column until an unvisited pixel is found that is above the previously determined threshold. Call this pixel P, and create an image object called O. (3) Mark P as visited and add P to O. (4) Consider the neighbors (pixels within one row or column) of P. For each neighbor that is above the threshold, call the neighbor P and repeat from step three. (5) Once steps 3 and 4 have concluded, O is a complete object in the image. Repeat from step 2. (6) Once all the pixels are visited, the O with the largest amount of pixels is considered the worm. (7) Set all pixels not in the largest O to 0. Relative animal size is then calculated as the number of pixels contained within its boundaries. Absolute size is obtained using scaling information contained in image metadata, or passed as an additional parameter. For cell-based studies, in which there may be multiple objects of interest in the same image, an optional size parameter can be set so that every object exceeding this value is identified.

2.3. A self-parameterizing thresholding algorithm

Once the locations, boundaries, and size of each animal are determined, the next task is to count H3-P-stained nuclei. Once absolute numbers are obtained, these can be converted to densities by dividing by animal size, effectively controlling for differences in animal sizes. The fact that stained neoblasts are often present at different pixel intensities complicates this operation, as there is no single threshold value that could be applied to the entire image that could accurately isolate the neoblasts. For these reasons, we implemented two different adaptive thresholding algorithms for inclusion in ImagePlane, a scale-dependent and scale-independent method. Complete details of these algorithms are given by Blayvas et al. (2006). In our hands, the adaptive thresholding methods performed best. This algorithm computes the threshold value for each pixel in the image through local weighted averages that are derived from max–min calculations across the interior of the animal in each image (Blayvas et al., 2006) (see Figs. 1 and 2 for additional details).

2.4. Image sectorization

Adapting previous work in *C. elegans* (Peng et al., 2008), we implemented a graph-based algorithm to automatically find the midline and left and right sides of an animal (see Fig. 3 for an example). Interestingly, we found that this algorithm performed poorly on some of our images. Further analyses determined that its performance was inversely proportional to the animals overall eccentricity; for example, it does well on long, thin animals, but poorly on more oval-shaped animals—a finding consistent with the

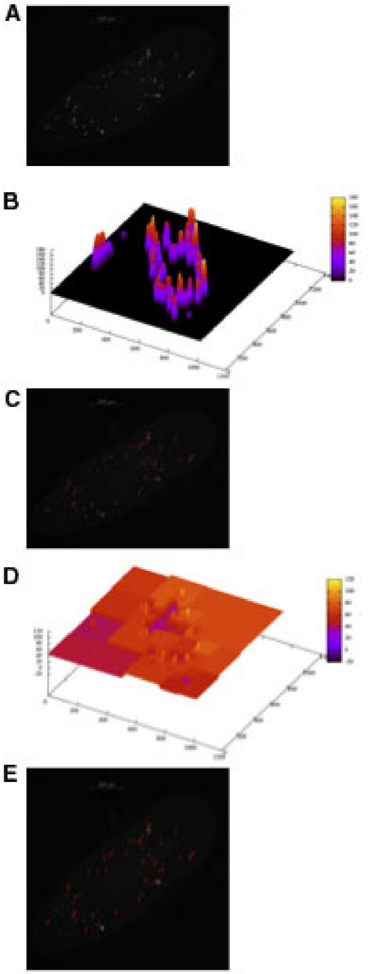


FIG. 1. Operation of the Niblac (scale dependent) and multiresolution (scale independent) automatic thresholding algorithms. These algorithms compute a thresholding surface that is used to isolate stained neoblasts. **(A)** A typical image of stained neoblasts in a flatworm. **(B)** The threshold surface computed by the Niblac algorithm for the image in **(A)**. **(C)** The automatically identified neoblasts (in red) after applying the computed threshold surface shown in **(B)**. **(D)** The threshold surface computed by the multiresolution algorithm for the image in **(A)**. **(E)** The automatically identified neoblasts (in red) after applying the computed threshold surface shown in **(D)**.

algorithm of Peng et al. (2008), since it was developed for processing images of *C. elegans*, which are long and thin, whereas planarians are more oval-shaped. Supplementary Figure S1 (Supplementary Material is available online at www.liebertonline.com/cmb) documents this phenomenon. It also demonstrates that our algorithm, which is based on a segmentation approach, performs much better on oval-shaped animals—the vast majority of planarian images (Fig. 3 and Supplementary Fig. S1). Asymmetric expression can be

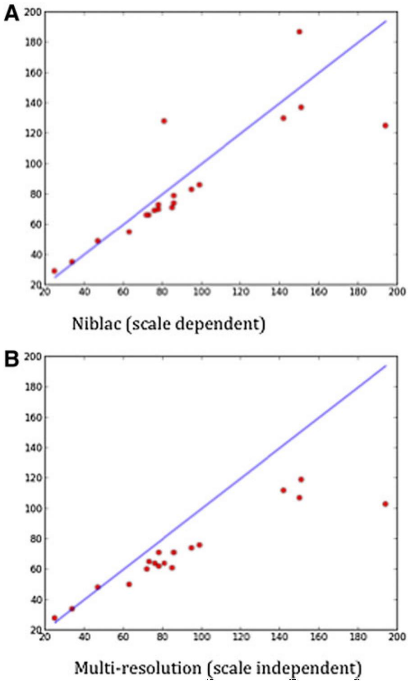


FIG. 2. Performance comparisons of Niblac (scale dependent) and multiresolution (scale independent) thresholding algorithms. X-axis, manual counts; Y-axis, automated counts. **(A)** Scale dependent. **(B)** Scale independent. The Pearson correlation coefficients are 0.94 for both **(A)** and **(B)**. Although the scale-dependent algorithm **(A)** is more accurate, it suffers from the requirement that users must select a threshold value, whereas the multiresolution scale-independent algorithm **(B)**, although less accurate, has the advantage of requiring no user configuration. ImagePlane supports both methods.

quantified by taking the ratio between the numbers of stained nuclei in an animal's A-P halves and L-R sides—note that even in the absence of knowledge of which end of an animal is anterior or which side is left, asymmetric expression along the A-P and L-R axes can still be measured and compared between sets of images. The statistical significance of the asymmetries is evaluated by randomly permuting the *x*,*y* coordinates of fluorescing nuclei, and rescoring each quadrant 100 times. Asymmetries larger than any of those found in the 100 permuted images are judged statistically significant.

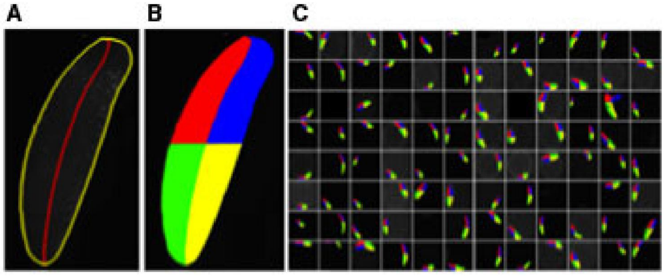
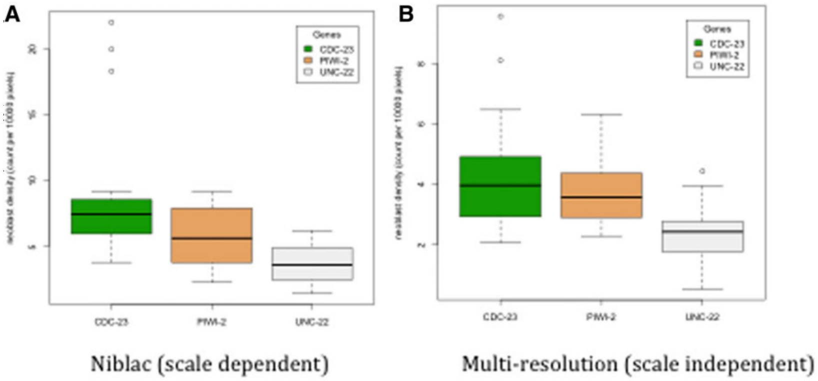


FIG. 3. Automated segmentation and sectorization of a 96-well plate. **(A)** Each worm's outline and midline are first determined. **(B)** The animals are then sectorized into four quadrants before counting cells (not shown). **(C)** A sample 96-well plate, automatically processed by ImagePlane *in situ*.



Niblac (scale dependent)

Multi-resolution (scale independent)

FIG. 4. Results summary for two different threshold algorithms. Summary of neoblast densities obtained automatically with ImagePlane for a dataset of 44 images with 18 *cdc23* knockdown animals, 11 *piwi2* knockdown animals, and a control set of 15 *unc22* animals. **(A)** Results using Niblac thresholding. Tukey adjusted *p*-values for comparisons between groups: *cdc23* to *piwi2* = 0.06, *cdc23* to *unc22* < 0.01, *piwi2* to *unc22* = 0.33. **(B)** Results using multi-resolution thresholding. Tukey adjusted *p*-values for comparisons between groups: *cdc23* to *piwi2* = 0.51, *cdc23* to *unc22* < 0.01, *piwi2* to *unc22* = 0.08. Error bars denote variance.

2.5. *Quantitation of results*

With methods in place to determine the size, location, and neoblast count of the animal in the image, the neoblast density is computed as the neoblast count divided by the size (in pixels) of the animal. These counts are output as simple tab-delimited files containing columns for image id, animal size, neoblast counts, and density. These files are easily imported into Excel and R for subsequent analyses and figure generation. Figure 4 shows an example output processed using an R macro (provided in software download) to generate a simple graphical report.

2.6. *Quantifying morphologies*

ImagePlane also includes an orientation and scale-invariant algorithm for detection, quantification, and analyses of morphological abnormalities such as those produced by RNAi gene knockdown experiments. The algorithm operates by first finding the outline of any automatically detected animal; this is accomplished using the algorithm previously described. ImagePlane then deduces the orientation of the animal by comparing the number of rows and columns transversed by the animal. Depending on the orientation of the animal, either the row or column lengths between the outline points are normalized by dividing by the size of the animal (in pixels) and recorded. Variation in the normalized distances produces a two-dimensional signature of the shape. Examples are shown in Figure 5A. These signatures have several important properties. First, they are orientation resistant, meaning that the orientation of the animal does not drastically affect the shape. Second, they are scale invariant; in other words, two animals with the same shape but of very different sizes will have similar signatures. Third, differences in any two signatures can be easily quantified, allowing us to group similar expression patterns and body morphologies (see Fig. 5B for an example).

3. RESULTS AND DISCUSSION

Planarians are renowned for their ability to regenerate (Reddien et al., 2005a); this ability is based on specialized neoblasts (Newmark and Sánchez Alvarado, 2000; Reddien et al., 2005b). These cells are the only proliferating cells in the planarian (Baguñà, 1974), and are found scattered throughout the animal. Neoblast progeny replace cells lost though normal cell turnover and are stimulated to proliferate when the animal is injured. With the successful introduction of dsRNA technology into planarians (Sánchez Alvarado and

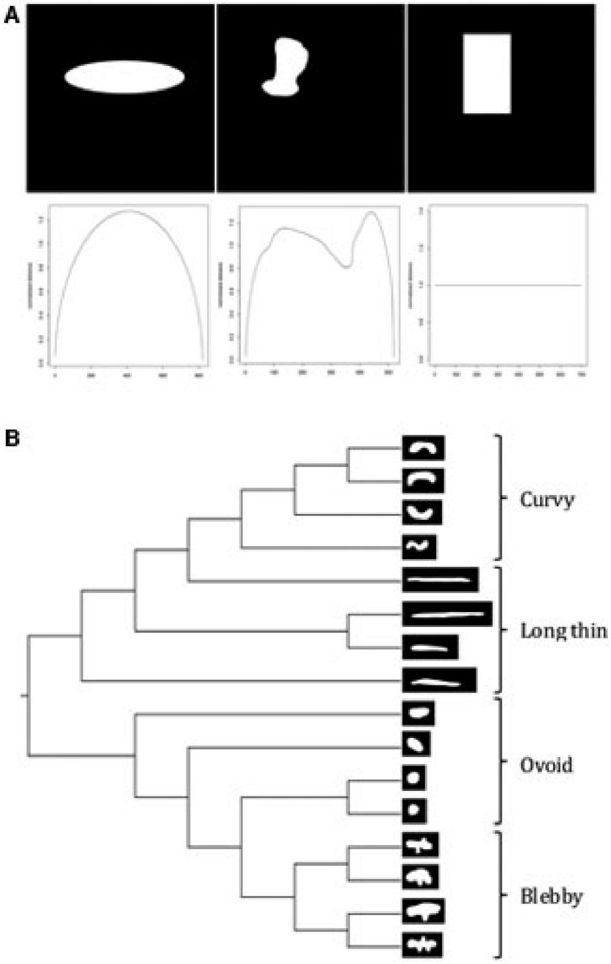


FIG. 5. A scale-invariant algorithm for quantifying morphologies. (**A**) Three randomly chosen shapes. Their corresponding morphological signatures are shown below each shape. (**B**) Cladogram representation of a neighbor-joining tree created using distances between the morphological signature for each shape shown on the tree's leaves.

Newmark, 1999; Reddien et al., 2005a), *S. mediterranea* has become the first invertebrate regeneration model system in which gene function can be analyzed. These facts, coupled with the availability of its annotated genome sequence (Cantarel et al., 2008; Robb et al., 2008), make *S. mediterranea* an ideal system to carry out functional genomics screens aimed at identification of genes involved in regulating stem cell proliferation. *S. mediterranea* has another equally important characteristic: it is literally a flatworm. This makes it ideal for image-based screens. To date, however, no general-purpose image analysis pipeline has been available to assist with high-throughput analyses of planarian images. We have developed ImagePlane to fill this need.

Automated analyses typically begin with automatic binarization of 96-well plate images, such as those generated by an automated confocal microscope, such at the BD imager. This step identifies the animal's outline and calculates its total pixel area. Next, an adaptive thresholding algorithm (Blayvas et al., 2006) is used to distinguish signal from noise within the interior of each animal's outline. This is necessary because the intensity of stained cells and nuclei differ from image to image and even within individual images (Fig. 1). Thus, an adaptive approach is desirable because it allows this threshold to vary in a dynamic fashion within the animal boundaries and between individual images. Another advantage of adaptive approaches such as Niback's (Blayvas et al., 2006) is that it requires minimal user inputs in order to determine the optimal threshold. This is a significant advantage, as it circumvents the need for users to manually inspect each image and to laboriously determine an optimal threshold by trial and error, as they would using a GUI-based platform such as Metamorph, Volocity, and ImageJ. After thresholding, ImagePlane uses its automated cell/particle counting algorithm—the same one used to identify animal boundaries in the first step of the pipeline—to identify stained cells and nuclei within animal boundaries.

To assess the accuracy of each of these steps, we carried out a double-blind experiment in which animals were H3-P-stained for neoblasts, which were manually counted, and compared these results to those produced automatically by ImagePlane. As Figure 2 indicates, the accuracy of the automated approach is very good as judged by a Pearson's *R* of both the Niblack and multiresolution methods. The Pearson correlation of both the Niblack and multiresolution methods is 0.94. A paired *t*-test was used to test for differences between the manual and automated counts. Niblack's method was not significantly different from the manual count, but the multiresolution was judged significantly different with a *p*-value <0.01. The high correlation but significant difference between the manual counting and the multiresolution method is because of moderate but consistent undercounting.

ImagePlane also provides automated means for identifying A-P and L-R inhomogeneity in neoblast densities, such as those that might be produced in RNAi knockdown experiments of morphogens governing cell proliferation (Reddien et al., 2005b). This is accomplished using a modified form of the algorithm developed by Peng et al. (2008) for *C. elegans* studies. ImagePlane's algorithm sectorizes animals into four quadrants: anterior, posterior, left, and right. This algorithm allows users to automatically identify differences in expression along the length of an animal and between its and L–R halves. Asymmetric expression is quantified by taking the ratio between the numbers of staining nuclei in each quadrant—note that even in the absence of knowledge of which end of an animal is anterior or which side is left, asymmetric expression along the A-P and L-R axes can still be measured and compared between sets of images. Figure 3 shows a sample 96-well plate for which each well's image has been automatically processed to identify the animal outlines and to sectorize them. Also provided is an automated means for determining the statistical significance of these asymmetries. This is done by randomly permuting the *x*,*y* coordinates of fluorescing nuclei, and rescoring each quadrant 100 times. Asymmetries larger than any of those found in the 100 permuted images are judged statistically significant. Although our results did not contain any such asymmetries for the genes we analyzed, this functionality will likely prove useful for those carrying out screens aimed at identification of asymmetrically localized transcripts and proteins.

Figure 4 summarizes a proof-of-principle analysis. This figure demonstrates the automated detection of the effects of RNAi knockdowns of two genes, *piwi2* and *cdc23*, known to be involved in planarian neoblast maintenance, and proliferation (Reddien et al., 2005a, 2005b); these are compared with a negative control, *unc22* (Moerman et al., 1986; Yandell et al., 1994; Reddien et al., 2005a) (see Methods). Previous experiments have shown that when the gene *cdc23* is knocked down (silenced) in *S. mediterranea*, the neoblast density increases as compared with wild-type animals (Reddien et al., 2005a). It has also been shown that when the gene *piwi2* is knocked down, the neoblast density remains the same as compared with a wild-type animal, but that progeny cells fail to divide (Reddien et al., 2005b). For our proof-of-principle analyses, a dataset of 44 images was collected with 18 animals fed RNAi knockdown constructs for *cdc23*, 11 animals for *piwi2*, and a control set of 15 animals that were treated to knock down *unc22*. All 44 images were analyzed using ImagePlane's scale-dependent and scale-independent algorithms (Fig. 4A and B, respectively), and for each image, the neoblast density was computed 7 days post-feeding of the RNAi construct. These results demonstrate that ImagePlane, using either algorithm, was able to automatically detect a significant difference between *unc22* and *cdc23* ($p < 0.05$)—with no difference between *piwi2* and *unc22* ($p > 0.05$), Tukey multiple comparison method (R Core Team, 2012).

ImagePlane also addresses another challenge frequently encountered in high-throughput image-based screens of whole cells and animals: the need to automatically detect and quantify morphological changes.

To speed such analyses, we have developed a scale-invariant algorithm that can automatically quantitate changes in body morphology. Although techniques currently exist to detect and quantify particular morphological changes, to our knowledge, our algorithm is the first to do so in an entirely *ab initio* fashion, and should be widely applicable to many different types of high-throughput biological screening assays. The algorithm operates by creating a two-dimensional summary or signature of each animal's outline. Examples are shown in Figure 5A. These signatures have several important properties. First, they are orientation resistant, meaning that the animal's orientation does not alter the signature appreciably. Second, the signatures are scale invariant; in other words, two animals with the same shape but of different sizes will have the same signature. This is a very desirable property for biological applications, especially for planarian research, as planarians differ quite dramatically in size from animal to animal and between experimental batches. A third advantage of this approach is that differences in any two signatures can be easily quantified. This makes it possible to rapidly and automatically group images with similar body morphologies. Figure 5B demonstrates this functionality, showing a neighbor-joining tree (Saitou and Nei, 1987) based on the signatures produced for images shown on the leaves.

3.1. Implementation and availability

ImagePlane is written in the Python programming language. ImagePlane consists of five basic Python modules that provide a set of interlocked methods that cover the essential activities that typify these screens: automatic determination of animal outlines and size; automatic image thresholding; methods for counting labeled populations of cells; and sectorization for quantitation of morphological changes induced by experimental manipulations. It is free for academic use and is available for download.

4. CONCLUSIONS

ImagePlane provides a set of interlocked methods that cover the essential activities of automatic determination of animal outlines and size; automatic image thresholding; methods for counting labeled populations of cells; and sectorization and quantitation of morphological changes induced by experimental manipulations. In the tradition of genome annotation pipelines, our goal has been to produce a practical pipeline for automated analysis of large collections of images, rather than to advance the basic science of image processing. As such, ImagePlane is an example of the new and growing domain of Bioimage informatics (Eliceiri et al., 2012) and is designed to support high-throughput 96-well screens such as the one described here (see Methods). Our goal has been to enable analyses of large numbers of images in an entirely automated fashion, without having to inspect a single image, and without extensive training or pipeline tuning procedures. As our results demonstrate, ImagePlane can analyze large numbers of images rapidly, accurately, and in an *ab initio* fashion.

ACKNOWLEDGMENTS

The authors would like to thank J. Jenkin, S. Robb, and J. Robb for their help and advice regarding experimental techniques; J. Pellettieri for help counting worms; Deepak Anthony for preliminary work on the project; and A. Sánchez Alvarado for help and encouragement throughout. This work was supported by R01GM088269 to M.Y.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Baguñà, J. 1974. Dramatic mitotic response in planarians after feeding, and a hypothesis for the control mechanism. *J. Exp. Zool.* 190, 117–122.

Blayvas, I., Bruckstein, A., and Kimmel, R. 2006. Efficient computation of adaptive threshold surfaces for image binarization. *Pattern Recognit*. 39, 89–101.

Boutros, M., Kiger, A.A., Armknecht, S., et al. 2004. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* 303, 832–835.

Cantarel, B.L., Korf, I., Robb, S.M.C., et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 18, 188–196.

Collins, T.J. 2007. ImageJ for microscopy. *BioTechniques* 43, 25–30.

Curwen, V., Eyras, E., Andrews, T.D., et al. 2004. The Ensembl automatic gene annotation system. *Genome Res*. 14, 942–950.

Eliceiri, K.W., Berthold, M.R., Goldberg, I.G., et al. 2012. Biological imaging software tools. *Nat. Methods*. 9, 697–710.

Fleischmann, R.D., Adams, M.D., White, O., et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512.

Hendzel, M.J., Wei, Y., Mancini, M.A., et al. 1997. Mitosis-specific phosphorylation of histone H3 initiates primarily within pericentromeric heterochromatin during G2 and spreads in an ordered fashion coincident with mitotic chromosome condensation. *Chromosoma* 106, 348–360.

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* 12, 491.

Kamath, R.S., and Ahringer, J. 2003. Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods* 30, 313–321.

Kuttenkeuler, D., and Boutros, M. 2004. Genome-wide RNAi as a route to gene function in *Drosophila*. *Brief Funct. Genomic. Proteomic*. 3, 168–176.

Liang, C., Mao, L., Ware, D., et al. 2009. Evidence-based gene predictions in plant genomes. *Genome Res*. 19, 1912–1923.

Moerman, D.G., Benian, G.M., and Waterston, R.H. 1986. Molecular cloning of the muscle gene unc-22 in *Caenorhabditis elegans* by Tc1 transposon tagging. *Proc. Natl. Acad. Sci. U. S. A*. 83, 2579–2583.

Newmark, P.A., and Sánchez Alvarado, A. 2000. Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev. Biol*. 220, 142–153.

Oliver, S.G., Van der Aart, Q.J., Agostoni-Carbone, M.L., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357, 38–46.

Paddison, P.J., and Hannon, G.J. 2003. siRNAs and shRNAs: skeleton keys to the human genome. *Curr. Opin. Mol. Ther*. 5, 217–224.

Papadopulos, F., Spinelli, M., Valente, S., et al. 2007. Common tasks in microscopic and ultrastructural image analysis using ImageJ. *Ultrastruct. Pathol*. 31, 401–407.

Peng, H., Long, F., Liu, X., et al. 2008. Straightening *Caenorhabditis elegans* images. *Bioinformatics* 24, 234–242.

Reddien, P.W., Bermange, A.L., Murfitt, K.J., et al. 2005a. Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. *Dev. Cell*. 8, 635–649.

Reddien, P.W., Oviedo, N.J., Jennings, J.R., et al. 2005b. SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* 310, 1327–1330.

Robb, S.M.C., Ross, E., and Sánchez Alvarado, A. 2008. SmedGD: the *Schmidtea mediterranea* genome database. *Nucleic Acids Res*. 36, D599–D606.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 4, 406–425.

Sánchez Alvarado, A., and Newmark, P.A. 1999. Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc. Natl. Acad. Sci. U. S. A*. 96, 5049–5054.

Yandell, M.D., Edgar, L.G., and Wood, W.B. 1994. Trimethylpsoralen induces small deletion mutations in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U. S. A*. 91, 1381–1385.

Address correspondence to:
Dr. Mark Yandell
Eccles Institute of Human Genetics
University of Utah
15 North 2030 East, Room 2100
Salt Lake City, UT 84112-5330

E-mail: myandell@genetics.utah.edu

CHAPTER 3

TRANSIENTLY ACTIVE WNT/Β-CATENIN SIGNALING IS NOT REQUIRED BUT MUST BE SILENCED FOR STEM CELL FUNCTION DURING MUSCLE REGENERATION

Published as: Murphy, M. M., Keefe, A. C., Lawson, J. A., Flygare, S. D., Yandell, M., & Kardon, G. (2014). Transiently Active Wnt/β-Catenin Signaling Is Not Required but Must Be Silenced for Stem Cell Function during Muscle Regeneration. *Stem Cell Reports*, *3*(3), 475–488. doi:10.1016/j.stemcr.2014.06.019

Contributions: I wrote software that was used in a significant part of the analysis and helped with statistical matters in the data analysis.

Stem Cell Reports
Article
ISSCR
OPEN ACCESS

Transiently Active Wnt/β-Catenin Signaling Is Not Required but Must Be Silenced for Stem Cell Function during Muscle Regeneration

Malea M. Murphy,[1,2,3] Alexandra C. Keefe,[1,2] Jennifer A. Lawson,[1] Steven D. Flygare,[1] Mark Yandell,[1] and Gabrielle Kardon[1,*]
[1]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA
[2]Co-first author
[3]Present address: Children's Research Institute at UT Southwestern, Dallas, TX 75390, USA
*Correspondence: gkardon@genetics.utah.edu

http://dx.doi.org/10.1016/j.stemcr.2014.06.019

SUMMARY

Adult muscle's exceptional capacity for regeneration is mediated by muscle stem cells, termed satellite cells. As with many stem cells, Wnt/β-catenin signaling has been proposed to be critical in satellite cells during regeneration. Using new genetic reagents, we explicitly test in vivo whether Wnt/β-catenin signaling is necessary and sufficient within satellite cells and their derivatives for regeneration. We find that signaling is transiently active in transit-amplifying myoblasts, but is not required for regeneration or satellite cell self-renewal. Instead, downregulation of transiently activated β-catenin is important to limit the regenerative response, as continuous regeneration is deleterious. Wnt/β-catenin activation in adult satellite cells may simply be a vestige of their developmental lineage, in which β-catenin signaling is critical for fetal myogenesis. In the adult, surprisingly, we show that it is not activation but rather silencing of Wnt/β-catenin signaling that is important for muscle regeneration.

INTRODUCTION

Adult vertebrate muscle has an exceptional capacity for regeneration, mediated by a dedicated population of muscle stem cells. These muscle stem cells, termed satellite cells, were first identified by their unique anatomical position between the sarcolemma and basement membrane of myofibers (Mauro, 1961). Subsequently, satellite cells were found to express the transcription factor Pax7 (Seale et al., 2000), and Pax7 is required for their maintenance in adult mice (Günther et al., 2013; Kuang et al., 2006; Oustanina et al., 2004; Relaix et al., 2006; von Maltzahn et al., 2013). Recent genetic labeling and ablation studies in mouse, using *Pax7^{CreER}* mice, have definitively established that satellite cells are the endogenous stem cells necessary and sufficient for muscle regeneration (Lepper et al., 2009, 2011; Murphy et al., 2011; Sambasivan et al., 2011). During regeneration, satellite cells activate, proliferate, and give rise to transit-amplifying myoblasts, which differentiate into myocytes that fuse with one another to form multinucleate myofibers. In addition, like other stem cells, satellite cells self-renew.

Canonical Wnt/β-catenin signaling is an important regulator of many adult stem cells (Holland et al., 2013) and has been proposed to be critical for satellite cells and muscle regeneration. Wnts are secreted glycoproteins that function as ligands, and β-catenin is the central mediator of canonical Wnt signaling (Niehrs, 2012). In the absence of Wnts, β-catenin is phosphorylated and targeted for degradation. The binding of Wnts to their receptors leads to the formation of stabilized, unphosphorylated β-catenin that translocates to the nucleus, where it binds to TCF/LEF proteins and activates transcription of Wnt-responsive genes. Many studies have identified Wnt pathway components as being active during muscle regeneration (Brack et al., 2008, 2009; Le Grand et al., 2009; Polesskaya et al., 2003; Zhao and Hoffman, 2004). Based largely on gain-of-function, primarily in vitro experiments, multiple labs have proposed that Wnt/β-catenin signaling is essential for muscle regeneration, although the conclusions of these papers are often contradictory (reviewed in von Maltzahn et al., 2012). However, no studies have explicitly examined in vivo whether Wnt/β-catenin signaling is necessary and sufficient specifically within satellite cells and their derivatives for muscle regeneration.

In this study, we use a highly sensitive reporter of Wnt/β-catenin signaling (*TCF/Lef:H2B-GFP^{Tg}*; Ferrer-Vaquer et al., 2010), as well as a reagent (*Pax7^{CreERT2}*) that our lab has generated to genetically manipulate satellite cells with high specificity and efficiency (Murphy et al., 2011), to test the role of this signaling pathway specifically within satellite cells and their derivatives during muscle regeneration. We find that Wnt/β-catenin signaling is transiently active in myoblasts during regeneration. However, β-catenin is not required cell autonomously for muscle regeneration. Instead, downregulation of transiently activated β-catenin is critical for limiting the regenerative response, as continuous regeneration deleteriously leads to increased fibrosis and an increased number of small myofibers. Thus, surprisingly, we show that it is not activation of Wnt/β-catenin signaling but rather silencing of this activation that is important for muscle regeneration.

CrossMark

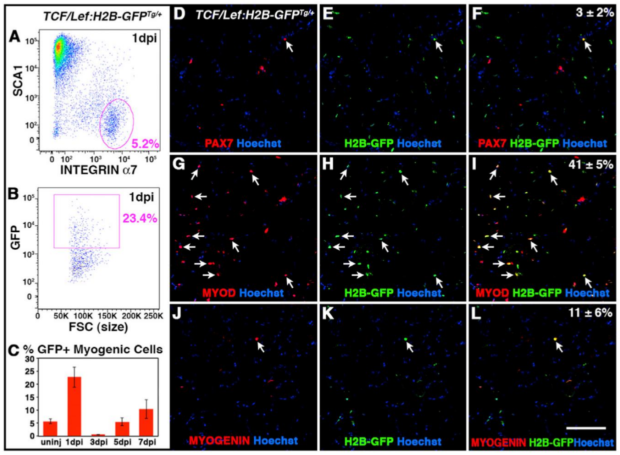


Figure 1. Wnt/β-Catenin Signaling Is Transiently Active in Myoblasts after Injury

(A–C) Mononuclear myogenic cells (A) transiently express *TCF/Lef:H2B-GFP* reporter at 1 dpi (n = 3 mice for each time point; B and C).

(D–L) At 1 dpi, only 3% of satellite cells (D–F) but 41% of myoblasts (G–I) and 11% of myocytes (J–L) are GFP+ (n = 3 mice). Arrows show GFP+ cells. The scale bar represents 100 μm.

Error bars in (C) represent one SEM.

RESULTS

Wnt/β-Catenin Signaling Is Transiently Active in Myoblasts during Regeneration

Multiple studies have established that Wnts, Frizzled receptors, nuclear β-catenin, coactivator BCL9, TCF/LEF reporters, and also Wnt antagonists secreted Frizzled-related proteins (sFRPs) are expressed during muscle regeneration (Brack et al., 2007, 2008; Le Grand et al., 2009; Otto et al., 2008; Polesskaya et al., 2003; Zhao and Hoffman, 2004). However, activation of Wnt/β-catenin signaling has not been explicitly tested and quantified within the myogenic lineage in vivo during the time course of regeneration. To test whether and when Wnt/β-catenin signaling is active in myogenic cells, we used the sensitive *TCF/Lef:H2B-GFP*^{Tg} reporter (Ferrer-Vaquer et al., 2010), in which cells with active Wnt/β-catenin signaling express nuclear localized GFP. To determine the percentage of myogenic cells with active Wnt/β-catenin signaling during regeneration, the right tibialis anterior (TA) muscles of *TCF/Lef:H2B-GFP*$^{Tg/+}$ mice were injured via $BaCl_2$ injection (Caldwell et al., 1990), injured TAs (and uninjured control TAs) collected at different days postinjury (dpi), and mononuclear myogenic cells analyzed via fluorescence-activated cell sorting (FACS). CD31-CD45-SCA1-INTEGRINα7+ cells were identified as myogenic (Yi and Rossi, 2011) and include satellite cells, myoblasts, and potentially myocytes (Figure 1A). In uninjured muscle, an average of 6% of myogenic cells was GFP+, indicating that Wnt/β-catenin is active in few myogenic cells (Figures 1B and 1C). However, at 1 dpi, 23% of myogenic cells were GFP+, although this declines to 0.6% by 3 dpi. To determine in which myogenic cells Wnt/β-catenin signaling is transiently active, we analyzed sections of TAs from *TCF/Lef:H2B-GFP*$^{Tg/+}$ mice at 1 dpi via immunofluorescence (Figures 1D–1L). Whereas only 3% of PAX7+ satellite cells and 11% of MYOGENIN+ myocytes were GFP+, 41% of MYOD+ cells were GFP+. MYOD+ cells may be either activated PAX7+MYOD+ satellite cells or PAX7−MYOD+ myoblasts. Because few PAX7+ cells were GFP+, we interpret the GFP+MYOD+ cells to be myoblasts. Thus we find that Wnt/β-catenin signaling is transiently active during muscle regeneration at 1 dpi, particularly in myoblasts.

Canonical Wnt/β-Catenin Signaling Is Effectively Abrogated in Satellite Cells and Their Progeny in *Pax7*$^{CreERT2/+}$;*β-Catenin* $^{Δ/fl2-6}$ Mice

Our analysis of *TCF/Lef:H2B-GFP*$^{Tg/+}$ mice demonstrates that Wnt/β-catenin signaling is transiently active in myoblasts during muscle regeneration. To test whether Wnt/β-catenin signaling is necessary specifically within myogenic cells for regeneration, we conditionally deleted *β-catenin* in satellite cells using *Pax7*$^{CreERT2/+}$;*β-catenin*$^{Δ/fl2-6}$;*Rosa*$^{mTmG/+}$ mice. In *Pax7*^{CreERT2} mice, Cre-mediated recombination occurs specifically and efficiently (>94% recombination) in *Pax7*+ satellite cells after delivery of tamoxifen (TAM) (Murphy et al., 2011). The *β-catenin* loss-of-function allele creates a functional null following Cre-mediated deletion of exons 2–6, thus inactivating signaling (Brault et al., 2001). The fate of recombined cells was tracked via the *Rosa*^{mTmG} reporter, which ubiquitously expresses membrane-bound *Tomato* until Cre-mediated recombination excises *Tomato*, resulting in membrane-bound *GFP* expression (Muzumdar et al., 2007). We analyzed *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ mice and compared them to *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/+}$;*Rosa*$^{mTmG/+}$ mice to control for any possible *β-catenin* heterozygous phenotype. Satellite cells are the only cells that express *Pax7* in uninjured muscle (Murphy et al., 2011). Therefore, by delivering TAM before injury, in control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/+}$;*Rosa*$^{mTmG/+}$ mice, nearly all satellite cells and their progeny express *GFP* and all are heterozygous for *β-catenin*, whereas in mutant *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ mice, nearly all satellite cells and their progeny express *GFP* and are null for *β-catenin*.

We tested whether *β-catenin* was efficiently and completely deleted in satellite cells and their progeny in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ mice. To test that exons 2–6 of *β-catenin* were genetically deleted, we isolated by FACS nonmyogenic TOMATO+ and myogenic GFP+ cells from TAs of control and mutant mice given five 10 mg doses of TAM, injured via $BaCl_2$, and harvested at 5 dpi (when there are maximal number of satellite cells; Murphy et al., 2011). We isolated genomic DNA and used PCR to identify *β-catenin wild-type* (*WT*), *fl2-6*, and *Δ2-6* alleles (Brault et al., 2001). In control mice, TOMATO+ and GFP+ cells were positive for both *WT* and *Δ2-6* alleles (Figure S1A available online). In contrast, in mutant mice, TOMATO+ cells contained both *fl2-6* and *Δ2-6* alleles, whereas myogenic GFP+ cells contained only the *Δ2-6* allele (Figure S1A). To test that genetic loss led to loss of β-catenin protein, we analyzed β-catenin protein expression in nonmyogenic TOMATO+ and myogenic GFP+ cells FACS isolated from TAs 5 dpi of control and mutant mice. Whereas β-catenin was detectable in 23% and 17% of nonmyogenic TOMATO+ cells in control and mutant mice, respectively, its detection in myogenic GFP+ cells dropped from 76% in control to 1% in mutant mice (Figures S1C and S1D). Together, these experiments indicate that *β-catenin* is effectively deleted in satellite cell-derived myogenic cells in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ mice in response to TAM.

We next tested whether Wnt/β-catenin signaling was effectively abrogated in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice. To do this, we generated control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{+/+}$;*TCF/Lef:H2B-GFP*$^{Tg/+}$ and mutant *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*TCF/Lef:H2B-GFP*$^{Tg/+}$ mice, gave them five TAM doses, injured TAs, and harvested muscle 1 dpi (when *TCF/Lef:H2B-GFP* reporter levels are highest in myogenic cells). CD31-CD45-SCA1-INTEGRINα7+ myogenic cells were isolated by FACS from control and mutant mice and analyzed for GFP. Whereas 25% of control myogenic cells were GFP+, only 2.5% of mutant myogenic cells were GFP+, indicating that, by 1 dpi, canonical Wnt/β-catenin signaling is nearly completely abolished (Figures S1E and S1F). To further determine whether *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice effectively abolished Wnt/β-catenin signaling in muscle, we analyzed these mice during development. We have previously shown, using *Pax7*$^{iCre/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice, that β-catenin regulates the number and slow fiber type of fetal myofibers (Hutcheson et al., 2009). If *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice work as effectively, fetal mice given TAM during development should demonstrate a similar phenotype. To test this, timed pregnant dams were given TAM (E11.5, E13.5, and E15.5), pups harvested at E18.5, and hind limbs sectioned and analyzed as before (Hutcheson et al., 2009). We found that, similar to *Pax7*$^{iCre/+}$;*β-catenin*$^{\Delta/fl2-6}$, in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice, there were fewer myofibers and a loss of slow myofibers in many muscles, particularly in the soleus (Figures S2A and S2B). Thus, *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$ mice effectively abrogate Wnt/β-catenin signaling in myogenic cells and recapitulate phenotypes previously reported for fetal myogenesis.

β-catenin Is Not Required for Satellite Cells to Regenerate Muscle or to Self-Renew

Having established that *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ mice, upon TAM delivery, effectively abrogate β-catenin signaling in myogenic cells, we tested whether satellite cells and their progeny require β-catenin to regenerate muscle. *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2-6}$;*Rosa*$^{mTmG/+}$ and littermate control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/+}$;*Rosa*$^{mTmG/+}$ mice were given five doses of TAM and then the right TA injured via $BaCl_2$. $BaCl_2$ injury induces a stereotyped pattern of muscle regeneration, with the peak of number of satellite cells and regenerating myofibers 5 dpi and regeneration complete by 28 dpi (Murphy et al., 2011). We found that, at 5 dpi, there was no difference between mutant and control muscle in either the number or proliferation of PAX7+ satellite cells (Figure 2A). Satellite cells give rise to MYOD+ cells, and the peak number of MYOD+ cells after $BaCl_2$ injury occurs at 3 dpi (Murphy et al., 2011; M.M.M. and G.K., unpublished data). Surprisingly, despite activation of the *TCF/Lef:H2B-GFP*$^{Tg/+}$ reporter in a large number of MYOD+ cells (see above), β-catenin deletion did not alter the number or proliferation of MYOD+ cells (Figure 2B). Myoblasts differentiate into myocytes, and these myocytes fuse into regenerating myofibers, characterized by their expression of embryonic myosin heavy chain (MyHCemb), an immature form of MyHC replaced by slow (MyHCI) and fast isoforms (MyHCII) as nascent myofibers mature. However, we found no difference in the amount of MyHCemb between mutant and control muscle at 5 dpi (Figure 2C), indicating loss of β-catenin does not affect regeneration of new myofibers. At 28 dpi, there continued to be no deleterious effect on muscle stem cells or regeneration. To determine whether β-catenin is required for satellite cells to self-renew and

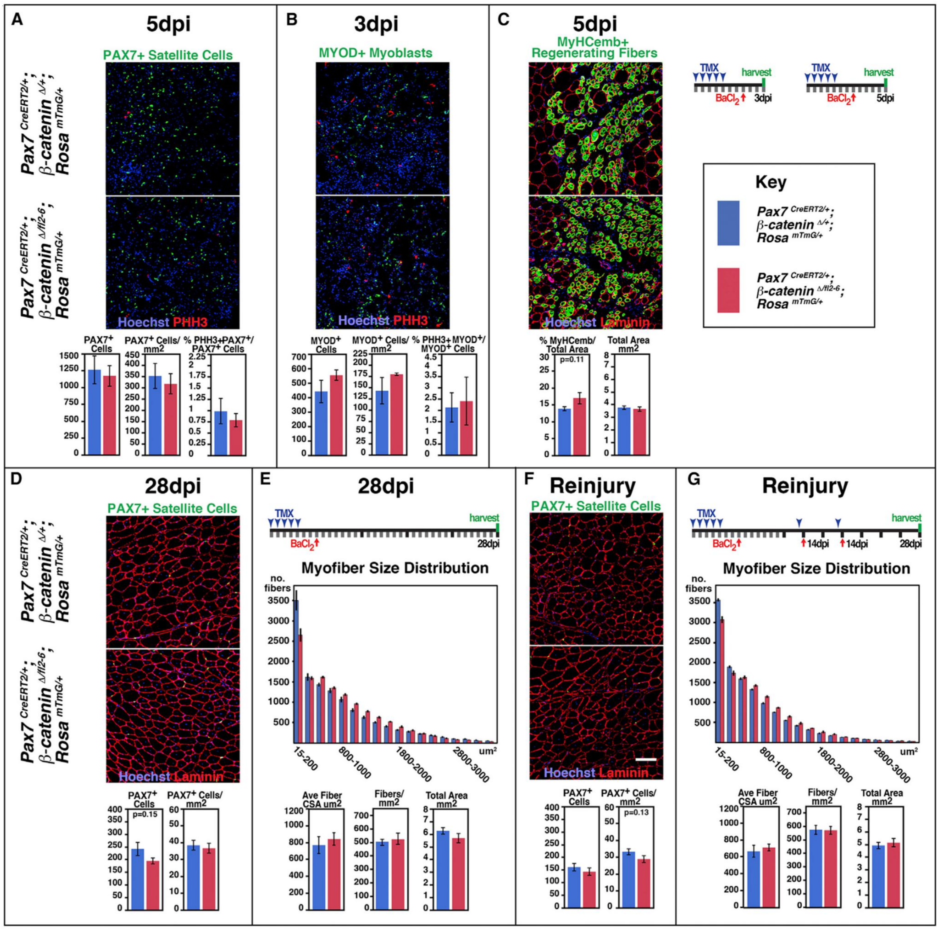


Figure 2. β-Catenin Is Not Required for Satellite Cells to Regenerate Muscle or Self-Renew

At 3 dpi (B; n = 3 control; n = 3 mutant), 5 dpi (A and C; n = 5 control; n = 5 mutant), and 28 dpi (D–G; n = 6 control; n = 5 mutant mice), β-catenin deletion does not affect number or proliferation of satellite cells (A and D), number or proliferation of myoblasts (B), amount of regenerating myofibers (C), or total muscle CSA (C and E). (F and G) After three rounds of injury (n = 5 control; n = 5 mutant mice), β-catenin deletion does not alter satellite cell self-renewal (F) or total muscle CSA (G). (E and G) At 28 dpi and after reinjury, myofibers are shifted to larger CSA when β-catenin is deleted. The scale bar (F) for all panels represents 100 μm. See also Figures S1–S3. Error bars in all histograms represent one SEM.

return to their niche, we compared the number of PAX7+ satellite cells between mutant and control mice at 28 dpi. However, neither the number nor location of satellite cells within their niche beneath the myofibers' LAMININ+ basal lamina differed (Figure 2D). In addition, regeneration was also unaffected, as neither the average myofiber cross-sectional area (CSA) nor the number of myofibers was affected by β-catenin deletion (Figure 2E). Interestingly,

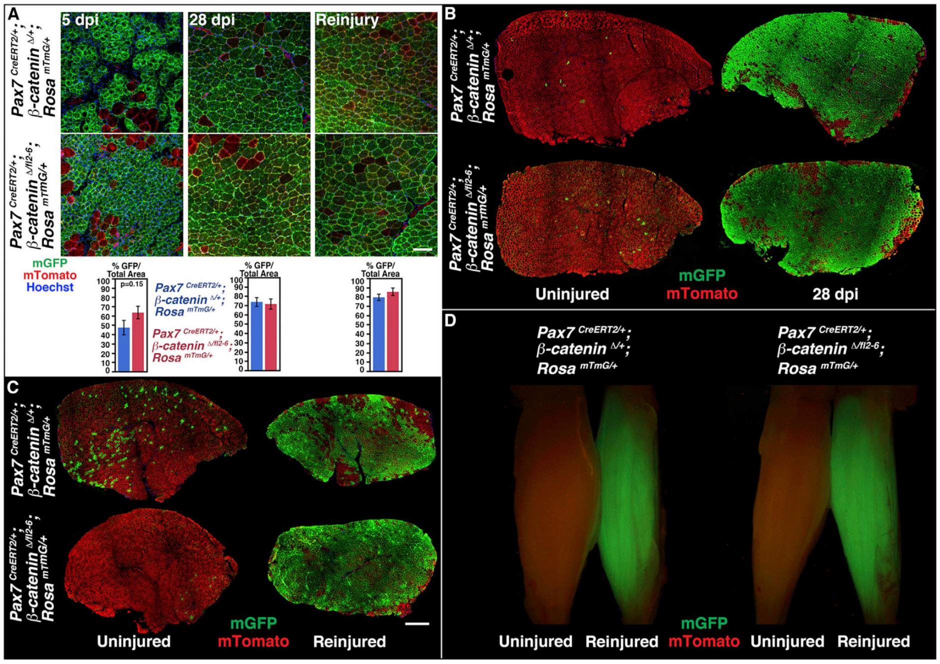

A
5 dpi
28 dpi
Reinjury
Pax7 CreERT2/+; β-catenin Δ/+; Rosa mTmG/+
Pax7 CreERT2/+; β-catenin Δ/fl2-6; Rosa mTmG/+
mGFP
mTomato
Hoechst
% GFP/ Total Area
p=0.15
100
90
80
70
60
50
40
30
20
10
0
Pax7 CreERT2/+; β-catenin Δ/+; Rosa mTmG/+
Pax7 CreERT2/+; β-catenin Δ/fl2-6; Rosa mTmG/+
B
Pax7 CreERT2/+; β-catenin Δ/+; Rosa mTmG/+
Pax7 CreERT2/+; β-catenin Δ/fl2-6; Rosa mTmG/+
Uninjured
mGFP
mTomato
28 dpi
C
Pax7 CreERT2/+; β-catenin Δ/+; Rosa mTmG/+
Pax7 CreERT2/+; β-catenin Δ/fl2-6; Rosa mTmG/+
Uninjured
mGFP
mTomato
Reinjured
D
Pax7 CreERT2/+; β-catenin Δ/+; Rosa mTmG/+
Pax7 CreERT2/+; β-catenin Δ/fl2-6; Rosa mTmG/+
Uninjured
Reinjured
mGFP
mTomato
Uninjured
Reinjured

Figure 3. Loss of β-Catenin Does Not Affect Satellite Cell Contribution to Regenerated Myofibers

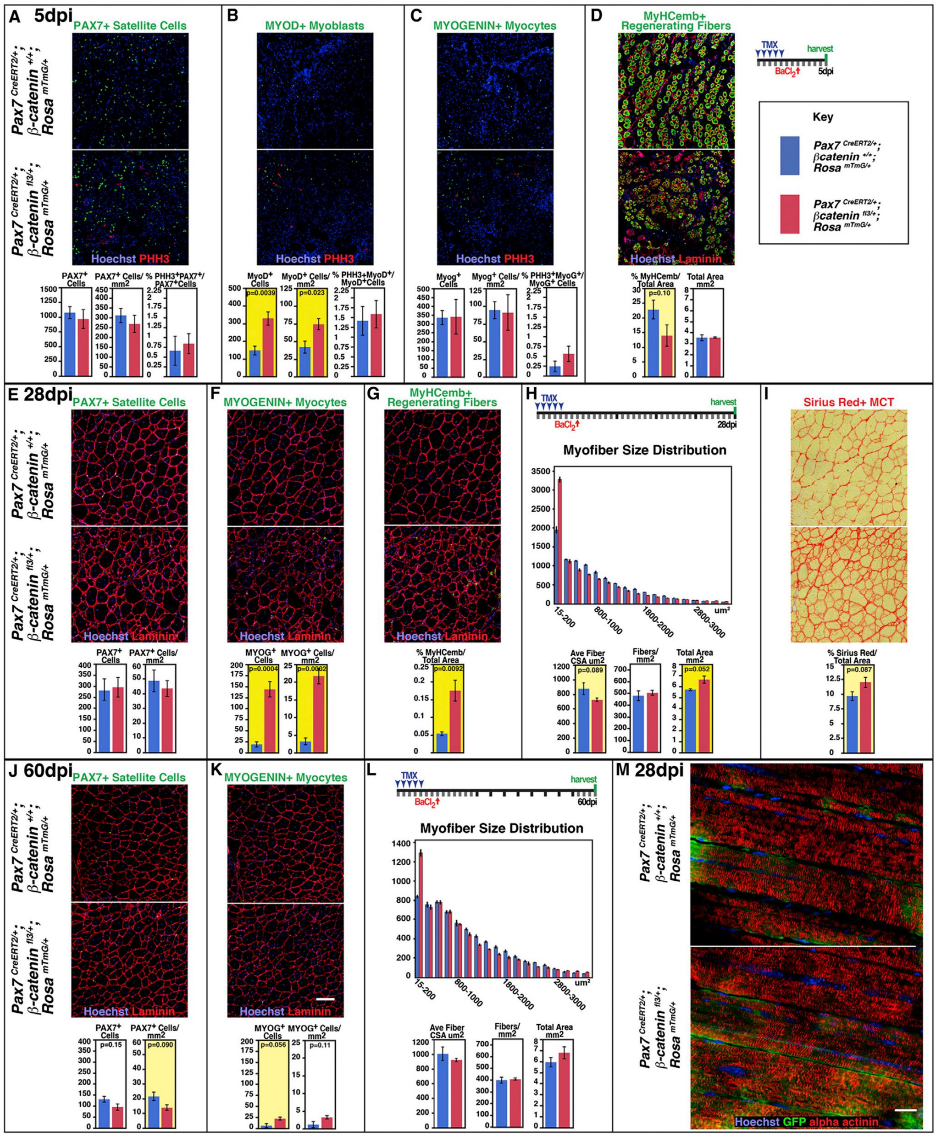
At 5 dpi (A; n = 5 control; n = 5 mutant), 28 dpi (A and B; n = 6 control; n = 5 mutant mice), or after reinjury (A, C, and D; n = 5 control; n = 5 mutant mice), β-catenin-null satellite cells regenerate GFP+ myofibers. Sections through entire contralateral and injured TA and extensor digotorum longus (EDL) muscles (B and C) and whole-mount images of contralateral and reinjured TAs (D). The scale bars represent 100 μm (A) or 0.5 mm (B and C). See also Figures S1–S3.

Error bars in (A) represent one SEM.

the distribution of the CSA of individual myofibers shifted to larger sizes with β-catenin deletion, but the overall area of the TA muscle was unaffected (Figure 2E). In summary, loss of β-catenin has no deleterious effect on the ability of satellite cells to self-renew, activate, proliferate, differentiate into myoblasts, or regenerate myofibers.

Although our data indicate that β-catenin is not required within myogenic cells for muscle regeneration, potentially the function of β-catenin may only be uncovered after multiple rounds of regeneration. To test this, we successively injured the TA and allowed it to regenerate three times (strategy in Figure 2G). Even after repeated rounds of regeneration, we detected no difference between mutant and control mice in satellite cell self-renewal, average myofiber CSA, or number of regenerated myofibers (Figures 2F and 2G). Similar to our findings at 28 dpi, we observed that the CSA of individual myofibers was shifted to larger sizes with loss of β-catenin, although the overall area of the TA was not changed (Figure 2G). Thus, β-catenin is not required in myogenic cells to regenerate muscle even after multiple rounds of regeneration.

A possible explanation for the lack of a defect in muscle regeneration with deletion of β-catenin may be technical issues with the *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2\text{-}6}$;*Rosa*$^{mTmG/+}$ mice. As satellite cells are highly proliferative, a few nonrecombined “escaper” satellite cells, retaining one wild-type allele of *β-catenin*, could potentially outcompete *β-catenin*-null cells and regenerate muscle. To test this, we compared the amount of GFP, representative of the contribution of satellite cells, in muscle from mutant and control mice and found no difference in GFP at 5 or 28 dpi or after multiple rounds of reinjury (Figure 3). Recently, it has been



(legend on next page)

shown that continuous administration of TAM during muscle regeneration may be required to completely delete a gene of interest in satellite cells (Günther et al., 2013). We repeated our experiments injuring TA muscles of mutant and control mice but with continuous TAM administration (strategy in Figure S3C) and analyzed muscles at 28 dpi. Similar to our previous results, we found no difference in satellite cell self-renewal or their contribution to regeneration (Figures S3A, S3B, and S3D). As we saw previously, the distribution of the CSA of individual myofibers was shifted to larger sizes with loss of β-catenin but now resulted in a slight increase (but not significant; p = 0.07) in the average CSA of myofibers and a slight decrease (but not significant; p = 0.08) in the number of myofibers, although the overall CSA of the muscle was unaffected (Figure S3C). Thus, we show in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2\text{-}6}$;*Rosa*$^{mTmG/+}$ mice that satellite cells are able, despite loss of β-catenin, to effectively regenerate muscle. Also, the finding that two different stringent TAM strategies give similar results argues that the lack of a deleterious phenotype is unlikely to be a false-negative result.

In summary, we show by in vivo conditional deletion of *β-catenin* that, despite activation of Wnt/β-catenin signaling within myogenic cells, β-catenin is not required within satellite cells or their derivatives for muscle regeneration or satellite cell self-renewal.

Constitutive Activation of β-Catenin in Satellite Cells Alters Myoblast Kinetics, Resulting in a Prolonged Regenerative Response

Our experiments demonstrate that, although β-catenin is not required, Wnt/β-catenin signaling is transiently active in myogenic cells during muscle regeneration. This presents an alternative hypothesis: whereas Wnt/β-catenin signaling is not required, once activated, prompt downregulation of signaling may be important for proper muscle regeneration. To test this, we constitutively activated β-catenin in satellite cells and their derivatives and assayed for effects on muscle regeneration.

To constitutively activate β-catenin, we used *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ mice. In the *β-catenin*^{fl3} allele, Cre mediates deletion of exon 3 and the formation of a stabilized, constitutively active form of β-catenin (Harada et al., 1999). We confirmed that GFP expression reflects recombination in the *β-catenin* locus by isolating by FACS GFP+ myogenic and TOMATO+ nonmyogenic cells from TA muscles of control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{+/+}$;*Rosa*$^{mTmG/+}$ and mutant *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ mice given five TAM doses, injured, and harvested 5 dpi. Using genomic DNA and PCR, we found in control mice both GFP+ and TOMATO+ cells contained only the *WT* allele (Figure S1B). In mutant mice, TOMATO+ cells contained both the *fl3* and *WT* alleles, whereas GFP+ cells had only the *WT* allele because the primer-binding sites for the *fl3* allele were deleted by recombination (Figure S1B). Therefore, after TAM delivery to *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ mice, GFP+ myogenic cells constitutively activate β-catenin.

We examined whether constitutive activation of β-catenin affected the expansion or self-renewal of satellite cells during muscle regeneration. Comparison of mutant *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ with littermate control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{+/+}$;*Rosa*$^{mTmG/+}$ mice revealed that, at 5 dpi, when the number of satellite cells peaks, there was no difference in the number or proliferation of PAX7+ satellite cells with constitutive β-catenin activation (Figure 4A). At 28 dpi, when muscle regeneration is complete, there was no difference in the number of satellite cells that had self-renewed (Figure 4E), although at 60 dpi, there was a slight (but not significant; p = 0.09) decrease in satellite cells with constitutive β-catenin activation (Figure 4J). Thus, constitutive β-catenin activation did not alter the expansion of satellite cells or significantly impair their return to the niche during regeneration.

The transient activation of Wnt/β-catenin signaling in myoblasts suggests that β-catenin may regulate myoblast expansion or differentiation during regeneration. Normally, the number of MYOD+ cells peaks at 3 dpi and declines by 5 dpi, and MYOD+ cells are absent at 28 dpi (Figures 2B and 4B; Murphy and Kardon, 2011; data not shown). At 5 dpi, there was a significant 1.79-fold increase in the number of MYOD+ cells (p = 0.02) with constitutive β-catenin activation, although there was no difference in proliferation of these cells (Figure 4B). There was no difference in either the number or proliferation of MYOGENIN+

Figure 4. Constitutive Activation of β-Catenin Alters the Kinetics of Myoblast Differentiation, Resulting in a Prolonged Regenerative Response

(A–D) At 5 dpi (n = 6 control; n = 6 mutant), constitutive β-catenin activation does not alter expansion or proliferation of satellite cells (A) but increases the number of myoblasts (B) at the expense of regenerating myofibers (D).

(E–I) At 28 dpi (n = 4 control; n = 5 mutant mice), β-catenin activation causes continued presence of myocytes (F) and regenerating myofibers (G), smaller regenerating myofibers (H), unresolved fibrosis (I), and results in TAs with larger CSA (H).

(J–L) At 60 dpi (n = 3 control; n = 3 mutant mice), with β-catenin activation, myocytes (K) and smaller myofibers (L) are still present.

(M) β-catenin activation does not alter sarcomere structure of regenerated myofibers (n = 3 control; n = 3 mutant mice).

The scale bar in (K) for sections (A–C), (E–G), and (I–K) represents 100 μm. The scale bar in (M) represents 10 μm. See also Figure S1. Error bars in all histograms represent one SEM.

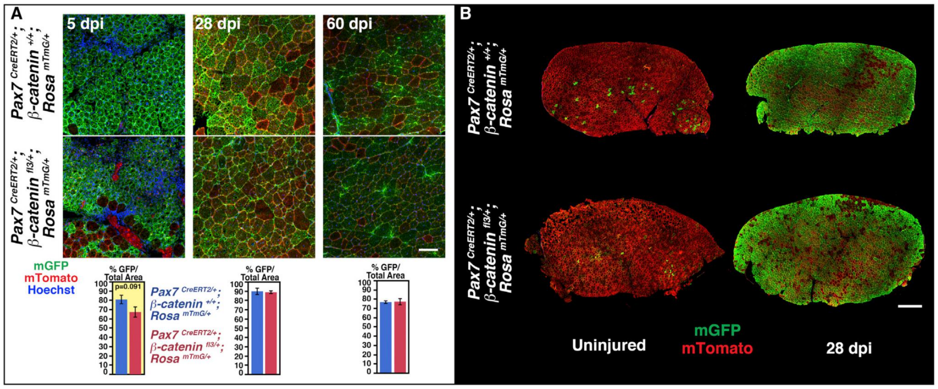


Figure 5. Constitutive Activation of β-Catenin Does Not Affect Satellite Cell Contribution to Regenerated Myofibers

At 5 dpi (A; n = 6 control; n = 6 mutant) or 28 dpi (A and B; n = 4 control; n = 5 mutant mice), satellite cells with constitutive β-catenin regenerate GFP+ myofibers. Sections through entire contralateral and injured TA and EDL muscles (B). The scale bars represent 100 μm (A) or 0.5 mm (B). See also Figure S1. Error bars in (A) represent one SEM.

myocytes (Figure 4C), but there was a slight decrease (high variance precluded significance; p = 0.10) in the amount of MyHCemb+-regenerating myofibers (Figure 4D). The lack of change in PAX7+ cells, increased number of MYOD+ cells, and decreased area of MyHCemb+-regenerating myofibers suggest that constitutive activation of β-catenin prolongs the time that myogenic cells remain as MYOD+ myoblasts.

To test if constitutive activation of β-catenin blocks myofiber differentiation, we examined TAs at 28 dpi, when regeneration is normally complete. No MYOD was expressed in either *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ or *Pax7*$^{CreERT2/+}$;*β-catenin*$^{+/+}$;*Rosa*$^{mTmG/+}$ mice (data not shown), and so the MYOD+ myoblasts present at 5 dpi do not remain in an undifferentiated state. In control mice, few MYOGENIN+ myocytes or actively regenerating MyHCemb+ myofibers remained at this time point, but in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ mice, there was a 6.8-fold increase in myocytes (p = 0.008) and a 3.6-fold increase in regenerating myofibers (p = 0.009; Figures 4F and 4G). This increase in regenerating myofibers was reflected in a shift to smaller myofibers (Figure 4H). Interestingly, the total CSA of the TAs was significantly increased with constitutive β-catenin activation (p = 0.05; Figure 4H). This increased muscle size may partially result from a slight increase in Sirius Red+ connective tissue (but not significant; p = 0.09; Figure 4I), and this increase in fibrosis may reflect that the regenerative response is ongoing. In total, these data show that, at 28 dpi, when muscle regeneration is normally complete, constitutive activation of β-catenin leads to a prolonged regenerative response, which is reflected in the continued presence of myocytes, actively regenerating myofibers and unresolved fibrosis.

We examined regenerated TAs at 60 dpi to test whether constitutive β-catenin activation had long-term effects on muscle regeneration. We found there was a 2.7-fold increase in MYOGENIN+ myocytes in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ compared with control mice (p = 0.05; Figure 4K). Although there was no expression of MyHCemb in either genotype (data not shown), there was an increased number of smaller (likely newly regenerated) myofibers (Figure 4L). Thus, even at 60 dpi, the regenerative response is ongoing.

Potentially, constitutive β-catenin activation could prevent myogenic cells from regenerating muscle, and all regenerated muscle could result from a small population of nonrecombined escaper cells. To test this, we compared GFP expression (which reflects constitutive β-catenin activation) in *Pax7*$^{CreERT2/+}$;*β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ and control *Pax7*$^{CreERT2/+}$;*β-catenin*$^{+/+}$;*Rosa*$^{mTmG/+}$ mice. At 5 dpi, there was slightly less GFP expression in mutant mice (but not significant; p = 0.09; Figure 5A), but this likely reflects the decrease in regenerated myofibers (Figure 4D). However, at 28 and 60 dpi, there was no difference in GFP expression between mutant and control mice, as nearly all regenerated myofibers with centralized myonuclei were GFP+ (Figures 5A and 5B). This demonstrates that, whereas constitutive activation of β-catenin prolongs the regenerative response, ultimately it does not prevent myogenic cells from regenerating muscle.

β-catenin is also a member of the adherens junction complex and localizes to the membrane of muscle fibers, and its overexpression can cause muscle structural defects (Kramerova et al., 2006; Nastasi et al., 2004). Thus, constitutive β-catenin activation in myogenic cells might cause muscle structural defects. To test this, we analyzed myofibers from $Pax7^{CreERT2/+};\beta\text{-}catenin^{fl3/+};Rosa^{mTmG/+}$ mice 4 weeks after injury. However, analysis of regenerated GFP+ myofibers for sarcomere structure did not reveal any obvious defects (Figure 4M).

In summary, constitutive activation of β-catenin in satellite cells and their derivatives prolongs the regenerative response to muscle injury. Whereas constitutive β-catenin activation does not prevent muscle regeneration or affect sarcomere structure of regenerated myofibers, it significantly affects the kinetics of muscle regeneration. Activation and proliferation of satellite cells is unaltered, but myoblasts and subsequently myocytes and smaller regenerating myofibers are present for an extended period, long after regeneration is normally complete.

β-Catenin Is Not Necessary but Is Sufficient to Regulate Slow Myofiber Type during Muscle Regeneration

β-catenin has been implicated in the determination of muscle fiber type. Previously, we showed that β-catenin is a necessary and sufficient positive regulator of the differentiation of slow MyHCI+ myofibers during fetal myogenesis (Hutcheson et al., 2009; Figure S2). Therefore, we tested in vivo whether β-catenin regulates the differentiation of slow myofibers during muscle regeneration. However, examination of mutant $Pax7^{CreERT2/+};\beta\text{-}catenin^{\Delta/fl2\text{-}6};Rosa^{mTmG/+}$ and control $Pax7^{CreERT2/+};\beta\text{-}catenin^{\Delta/+};Rosa^{mTmG/+}$ mice revealed no difference in MyHCI expression at either 5 or 28 dpi (Figures 6A and 6B). In satellite cell-derived C2C12 cells, expression of fast MyHCIIb is directly regulated by β-catenin binding, via TCF/LEF, to the MyHCIIb promoter (Shanely et al., 2009). Again, examination of mutant and control mice revealed no difference in MyHCIIb expression at 28 dpi (Figure 6C). Because constitutive β-catenin is sufficient to drive all fetal myogenic progenitors to differentiate into slow MyHCI+ myofibers (Hutcheson et al., 2009), we tested whether constitutive β-catenin activation in satellite cells would have a similar effect on regenerated myofibers. We found a 4.5-fold increase in MyHCI expression ($p = 0.0004$) in $Pax7^{CreERT2/+};\beta\text{-}catenin^{fl3/+};Rosa^{mTmG/+}$ versus $Pax7^{CreERT2/+};\beta\text{-}catenin^{+/+};Rosa^{mTmG/+}$ mice at 5 dpi (Figure 4D). At 28 dpi, there continued to be a 4-fold increase in MyHCI expression ($p = 0.009$; Figure 4E) with constitutive β-catenin activation, and this effect was somewhat maintained at 60 dpi (but not significant; $p = 0.09$; Figure 6F). However, whereas only a small percentage of the myofibers are MyHCI+, 80%–90% of the myofibers are GFP+ and satellite cell derived (Figure 5A). This small number of MyHCI+ myofibers may reflect an incomplete conversion to a slow fiber type or that nerve-derived signals significantly modulate fiber type. Altogether, our data demonstrate that, during adult muscle regeneration, β-catenin is not necessary for differentiation of slow MyHCI+ or fast MyHCIIb+ myofibers. However, β-catenin is sufficient to cell autonomously positively regulate MyHCI expression but, unlike during fetal myogenesis, cannot convert all myofibers to a slow MyHCI+ fiber type during regeneration.

DISCUSSION

Wnt/β-catenin signaling has been proposed to be critical for adult muscle regeneration (reviewed in von Maltzahn et al., 2012). Here, we explicitly test the role of this signaling pathway specifically within satellite cells and their derivatives during muscle regeneration in vivo. We find that Wnt/β-catenin signaling is transiently active in myoblasts during muscle regeneration. However, unlike previous studies, we find that β-catenin is not required in myogenic cells for regeneration, but instead downregulation of transiently activated β-catenin is critical for limiting the regenerative response. Thus, we show that it is not activation but rather silencing of Wnt/β-catenin signaling that is important for muscle regeneration (summarized in Figure 7).

Using the highly sensitive $TCF/Lef{:}H2B\text{-}GFP^{Tg}$ reporter, we demonstrate that Wnt/β-catenin signaling is transiently active during muscle regeneration, specifically in myoblasts 1 dpi. Our finding that myoblasts transiently transduce Wnt/β-catenin signals agrees with previous analyses of nuclear β-catenin and *TOPGAL* reporter expression (Brack et al., 2007, 2008). Results from others (Abiola et al., 2009; Le Grand et al., 2009; Polesskaya et al., 2003; Zhao and Hoffman, 2004) show that Wnts are upregulated during regeneration, although the cellular origin of these Wnts is unresolved. The Wnt antagonists, sFRPs 1, 2, and 4 are also strongly upregulated during regeneration (Le Grand et al., 2009; Zhao and Hoffman, 2004), and this likely is the endogenous molecular mechanism by which Wnt/β-catenin signaling, activated at 1 dpi, is subsequently silenced.

We explicitly tested the requirement for β-catenin in satellite cells and their derivatives for muscle regeneration. Despite efficient deletion of β-catenin, satellite cells were able to self-renew and regenerate muscle (although a subtle phenotype, undetectable in our assays, is possible). Interestingly, we did see that with β-catenin deletion myofibers shifted to larger cross-sectional areas at 28 dpi or with reinjury. Given that constitutive β-catenin activation prolonged regeneration and resulted in a shift to smaller, regenerating myofibers, the shift to larger myofibers with

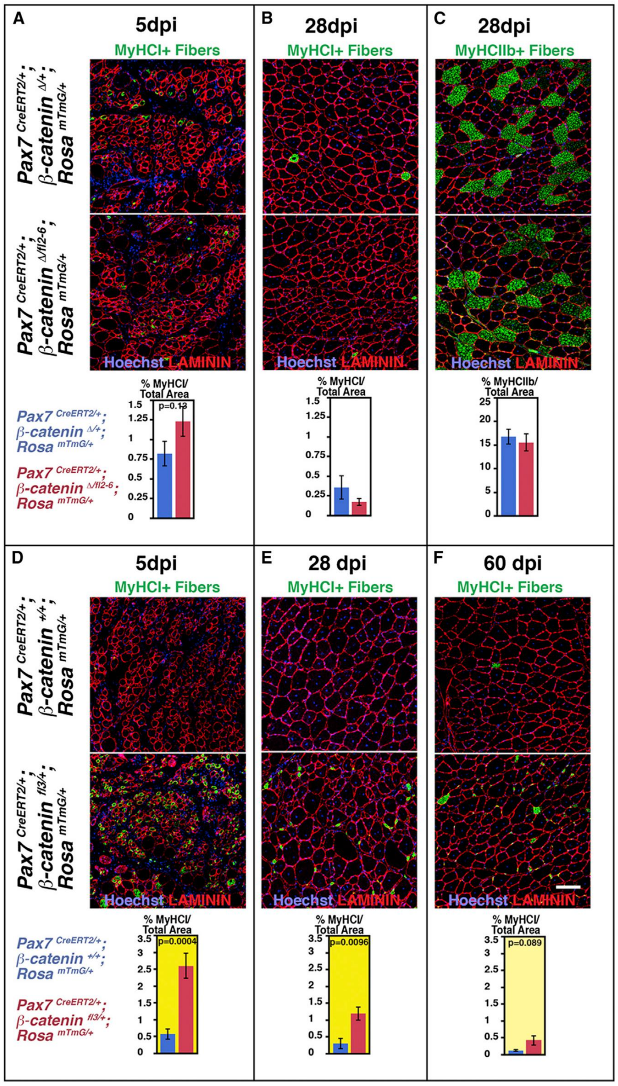


Figure 6. β-Catenin Is Not Necessary but Is Sufficient to Regulate Slow Myofibers during Regeneration

(A–C) β-catenin deletion does not alter the amount of slow MyHCI+ (A and B) or fast MyHCIIb (C) myofibers (at 5 dpi: n = 5 control, n = 5 mutant; at 28 dpi: n = 6 control, n = 5 mutant mice).

(D–F) β-catenin activation increases the amount of slow MyHCI+ myofibers (at 5 dpi: n = 6 control, n = 6 mutant; at 28 dpi: n = 4 control, n = 5 mutant).

Error bars in all histograms represent one SEM.

loss of β-catenin may indicate premature differentiation of myofibers. However, the finding that TAs regenerated from β-catenin satellite cells are GFP+ and do not differ in overall size from control TAs suggests that a potential requirement of β-catenin to inhibit premature differentiation is modest, at best. Consistent with the lack of a significant phenotype

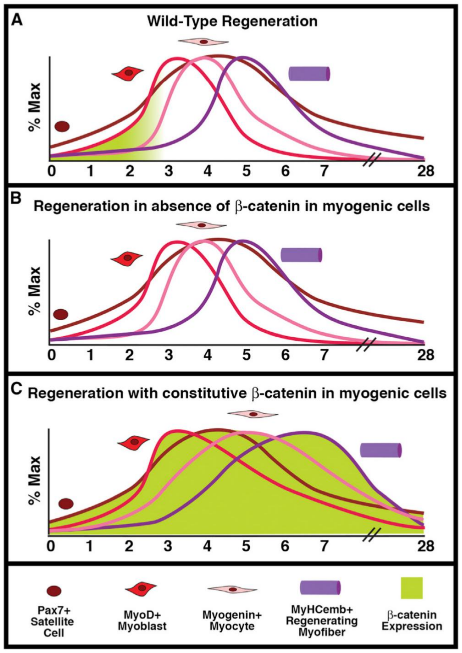


Figure 7. Model of Role of Wnt/β-Catenin Signaling in Adult Muscle Regeneration

(A) During wild-type regeneration, Wnt/β-catenin signaling is transiently active in myoblasts.

(B and C) Deletion of β-catenin in satellite cells and their derivatives does not alter muscle regeneration (B), but constitutive β-catenin activation alters the kinetics of myoblast differentiation (C), leading to a prolonged regenerative response.

with β-catenin loss, previous studies have produced contradictory findings. Most experiments have been conducted in vitro and, using a variety of techniques to inhibit β-catenin signaling, have found decreased satellite cell proliferation (Otto et al., 2008), less differentiation (Brack et al., 2008; Descamps et al., 2008; Kim et al., 2008), or more differentiation (Gavard et al., 2004; Tanaka et al., 2011). Only Brack et al. (2008, 2009) inhibited Wnt/β-catenin signaling in vivo, via injection of sFRPs into regenerating TAs or genetic deletion of β-catenin coactivators BCL9 and BCL9-2 (via *Myf5*$^{Cre/+}$;*Bcl9*$^{loxP/loxP}$;*Bcl9-2*$^{loxP/loxP}$ mice) after $BaCl_2$ or freeze injury. They concluded that Wnt/β-catenin is necessary to promote muscle differentiation, but addition of sFRPs blocks both canonical and noncanonical Wnt signaling (Li et al., 2008) and does not specifically target myogenic cells, and the genetic BCL9/BCL9-2 deletion potentially affects satellite cell development. Thus, previous phenotypes attributed to β-catenin necessity in satellite cells for regeneration may reflect in vitro conditions or in vivo reveal the role of canonical signaling in muscle progenitors during development or in other cell types involved in muscle regeneration or the function of noncanonical signaling in regeneration.

The transient activation of Wnt/β-catenin signaling in myoblasts suggested the alternative hypothesis that not activation but silencing of signaling is critical for proper muscle regeneration. To test this, we examined the effects of constitutive β-catenin activation. Previous studies testing this have primarily been conducted in vitro, via Wnt3a or LiCl delivery to cultured satellite cells, and found constitutive Wnt/β-catenin signaling either prevents differentiation (Gavard et al., 2004; Kuroda et al., 2013; Tanaka et al., 2011) or promotes differentiation and fusion (Bernardi et al., 2011; Brack et al., 2008; Han et al., 2011; Pansters et al., 2011). Two papers (Brack et al., 2008; Le Grand et al., 2009) tested in vivo (using either $BaCl_2$ or freeze injury) the effects of increased signaling and concluded that Wnt/β-catenin signaling promoted premature differentiation. However, these in vivo experiments activated, via ectopic Wnt3a, signaling in all cell types (including muscle connective tissue fibroblasts and endothelial cells) during regeneration. Our experiments constitutively activating β-catenin specifically in satellite cells revealed that satellite cells are largely insensitive to increased β-catenin, as we saw no effects on satellite cell expansion or proliferation after injury. However, constitutive β-catenin activation did alter the kinetics of the regenerative process, as myoblasts (which normally transiently express β-catenin) and subsequently myocytes and regenerating myofibers are present for an extended period. Thus, continued activation of β-catenin signaling prolongs the myoblast phase of regeneration, although it does not ultimately block differentiation. This prolonged regeneration negatively impacts muscle structure as it results in smaller myofibers and increased fibrosis.

Our study is a cautionary warning about the conclusions of Wnt/β-catenin signaling functional significance that can be drawn from reporter and gain-of-function experiments. Multiple reporters have been developed for Wnt/β-catenin signaling (reviewed in Barolo, 2006). Although they differ in their sensitivities, a finding of reporter activity is considered a good indicator that endogenous Wnt/β-catenin signaling is active. However, our data show that presence of activity does not necessarily imply a biological requirement for that activity. In addition, gain-of-function experiments reveal whether cells are sensitive to Wnt/β-catenin signaling but again do not demonstrate signaling necessity. Loss-of-function studies, particularly when conducted conditionally (to limit the spatial and temporal scope of deletion) and in vivo, are essential for determining the necessity of signaling. In fact, in our study, the combination of reporter and gain- and loss-of-functions experiments show that it is not activation but rather silencing of signaling that is important. In wild-type mice, this silencing is likely accomplished by the early and strong upregulation of sFRPs during muscle regeneration (Le Grand et al., 2009; Polesskaya et al., 2003; Zhao and Hoffman, 2004).

The finding that Wnt/β-catenin signaling must be silenced during adult regeneration naturally raises the question of why signaling is activated. Our previous analysis of Wnt/β-catenin signaling during fetal myogenesis demonstrated that β-catenin is required for regulation of fetal myofiber differentiation (Hutcheson et al., 2009). Similar to the classic argument about the origin and function of the Spandrels of San Marcos (Gould and Lewontin, 1979), activation of Wnt/β-catenin signaling in adult myoblasts might simply be a vestige of their developmental lineage, in which β-catenin signaling is required for fetal myogenesis (Hutcheson et al., 2009; Murphy and Kardon, 2011).

Comparison of the role of Wnt/β-catenin signaling during fetal myogenesis and adult regeneration reveals intriguing similarities and differences between fetal and adult stem cells and myoblasts. Loss-of-function experiments demonstrate that, during fetal myogenesis, β-catenin is critical for regulating the number and fiber type of myofibers that differentiate from PAX7+ stem cells and yet, in the adult β-catenin, is dispensable for the differentiation and fiber type of myofibers regenerated from PAX7+ satellite cells. This difference between fetal and adult stem cells is reminiscent of the difference between the development and maintenance of hematopoietic stem cells (HSCs); β-catenin is essential for HSC generation (Ruiz-Herguido et al., 2012) but later appears to be dispensable for maintenance of embryonic HSCs (Ruiz-Herguido et al., 2012) and adult HSC function (Cobas et al., 2004). Gain-of-function experiments demonstrate that differentiating fetal and adult myofibers are similarly sensitive to β-catenin, as constitutive β-catenin activation is sufficient to convert both types to slow MyHCI+ myofibers, although this conversion is more complete during fetal myogenesis. However, most striking is the difference in β-catenin sensitivity between fetal and adult PAX7+ stem cells. In the fetus, constitutive activation of β-catenin causes a dramatic expansion of PAX7+ stem cells (Hutcheson et al., 2009). In contrast, in the adult β-catenin, activation does not expand the number of PAX7+ satellite cells but rather the number of transit-amplifying MYOD+ myoblasts. Altogether, our experiments indicate that, despite their close lineage relationship (Hutcheson et al., 2009; Murphy and Kardon, 2011), fetal and adult PAX7+ stem cells differ in their requirement of and sensitivity to β-catenin. Limiting the sensitivity of highly proliferative satellite cells to β-catenin may be important for decreasing the adult risk of cancer from oncogenic β-catenin signaling.

EXPERIMENTAL PROCEDURES

Mice

All mouse lines were previously reported: *Pax7*^{CreERT2} (Murphy et al., 2011); *β-catenin*$^{fl2\text{-}6}$ (Brault et al., 2001); *β-catenin*^{fl3} (Harada et al., 1999); *Rosa*^{mTmG} (Muzumdar et al., 2007); and *TCF/Lef:H2B-GFP*^{Tg} (Ferrer-Vaquer et al., 2010). Mice were bred onto C57/BL6J background and used at 6–8 weeks of age.

FACS Cell Isolation and Analysis

Mononuclear myogenic cells and fibroblasts were isolated from injured right and uninjured left TAs, incubated with antibodies if needed (Table S1), and analyzed via FACS (details in Supplemental Experimental Procedures). Myogenic cells and fibroblasts were isolated from *TCF/Lef:H2B-GFP*^{Tg} mice using strategy of Yi and Rossi (2011). For *Pax7*$^{CreERT2/+}$;*β-catenin*$^{\Delta/fl2\text{-}6}$;*Rosa*$^{mTmG/+}$ and *Pax7*$^{CreERT2/+}$; *β-catenin*$^{fl3/+}$;*Rosa*$^{mTmG/+}$ mice and their controls, myogenic cells were isolated via GFP and nonmyogenic cells via TOMATO. Genomic DNA of cells was isolated and alleles of *β-catenin* determined via PCR using primers of Brault et al. (2001) and Harada et al. (1999).

Muscle Injury and Tamoxifen Delivery

Injury was induced by injecting 25 µl of 1.2% $BaCl_2$ in normal saline into right TA. Left TA served as uninjured control. For administration prior to injury, TAM was delivered via gavage in 10 mg doses. For continuous delivery before and after injury, TAM was delivered intraperitoneally at 3 mg/40 g body weight per injection. All mouse experiments were conducted under the oversight of University of Utah Institutional Animal Care and Use Committee.

Immunofluorescence, Histology, and Microscopy

For section immunofluorescence, flash-frozen muscles were sectioned, fixed in paraformaldehyde, and labeled via immunofluorescence or stained with Sirius Red (details in Supplemental Experimental Procedures). Sirius Red sections were imaged on a Zeiss Axioplan2 microscope. Immunofluorescent sections were imaged on a Nikon AR1 confocal or widefield microscope. Each confocal image is a composite of maximum projections, derived from stacks of optical sections.

Quantification and Statistics

Quantification of PAX7+, MYOD+, or MYOGENIN+ nuclei and amount of GFP, MyHCemb, MyHCI, MyHCIIb, or ECM was quantified using Image J (details in Supplemental Experimental Procedures). For each variable, counts of two sections across the entire TA were averaged for three to six individuals of each genotype per time point and analyzed using a Student's two-tailed t test. On all bar charts, mean ± 1 SEM shown. Fiber distribution was determined using MuscleQNT (developed by S.D.F. and M.Y. and available at https://github.com/stevendflygare/muscleQNT). In brief, MuscleQNT is an image analysis pipeline implemented in Python designed to identify borders (through adaptive thresholding and a series of erosion and dilation steps) of LAMININ+ myofibers and quantify the number and CSA of all myofibers in a muscle cross-section. Histograms and summary statistics of myofiber sizes are generated, and histogram error bars are the result of permutation tests. All displayed histograms were statistically significant via the Kolmogorov-Smirnov test.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.stemcr.2014.06.019.

AUTHOR CONTRIBUTIONS

M.M.M., A.C.K., J.A.L., and G.K. designed experiments; M.M.M., A.C.K., and J.A.L. performed experiments; S.D.F. and M.Y. designed and implemented MuscleQNT; M.M.M., A.C.K., J.A.L., and G.K. analyzed data; and M.M.M. and G.K. wrote the manuscript.

ACKNOWLEDGMENTS

We thank D. D. Cornelison, R. Dorsky, L. C. Murtaugh, and A. Sanchez-Alvarado for discussions and technical advice; C. Rodesch for microscopy support; J. Marvin and C. Leukel for flow cytometry support; and M. P. Colasanto, Z. Fox, A. Merrell, and L. C. Murtaugh for manuscript comments. Development of MuscleQNT by S.D.F. and M.Y. was supported by NIH R01GM088269 to M.Y. M.M.M. was supported by NIH Genetics Training Grant (T32 GM07464) and A.C.K. by NIH Hematology Grant (T32 DK007115). This work was supported by NIH R01 HD053728 and MDA 130903 to G.K.

Received: February 3, 2014
Revised: June 27, 2014
Accepted: June 30, 2014
Published: July 31, 2014

REFERENCES

Abiola, M., Favier, M., Christodoulou-Vafeiadou, E., Pichard, A.L., Martelly, I., and Guillet-Deniau, I. (2009). Activation of Wnt/beta-catenin signaling increases insulin sensitivity through a reciprocal regulation of Wnt10b and SREBP-1c in skeletal muscle cells. PLoS ONE *4*, e8509.

Barolo, S. (2006). Transgenic Wnt/TCF pathway reporters: all you need is Lef? Oncogene *25*, 7505–7511.

Bernardi, H., Gay, S., Fedon, Y., Vernus, B., Bonnieu, A., and Bacou, F. (2011). Wnt4 activates the canonical β-catenin pathway and regulates negatively myostatin: functional implication in myogenesis. Am. J. Physiol. Cell Physiol. *300*, C1122–C1138.

Brack, A.S., Conboy, M.J., Roy, S., Lee, M., Kuo, C.J., Keller, C., and Rando, T.A. (2007). Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis. Science *317*, 807–810.

Brack, A.S., Conboy, I.M., Conboy, M.J., Shen, J., and Rando, T.A. (2008). A temporal switch from notch to Wnt signaling in muscle stem cells is necessary for normal adult myogenesis. Cell Stem Cell *2*, 50–59.

Brack, A.S., Murphy-Seiler, F., Hanifi, J., Deka, J., Eyckerman, S., Keller, C., Aguet, M., and Rando, T.A. (2009). BCL9 is an essential component of canonical Wnt signaling that mediates the differentiation of myogenic progenitors during muscle regeneration. Dev. Biol. *335*, 93–105.

Brault, V., Moore, R., Kutsch, S., Ishibashi, M., Rowitch, D.H., McMahon, A.P., Sommer, L., Boussadia, O., and Kemler, R. (2001). Inactivation of the beta-catenin gene by Wnt1-Cre-mediated deletion results in dramatic brain malformation and failure of craniofacial development. Development *128*, 1253–1264.

Caldwell, C.J., Mattey, D.L., and Weller, R.O. (1990). Role of the basement membrane in the regeneration of skeletal muscle. Neuropathol. Appl. Neurobiol. *16*, 225–238.

Cobas, M., Wilson, A., Ernst, B., Mancini, S.J., MacDonald, H.R., Kemler, R., and Radtke, F. (2004). Beta-catenin is dispensable for hematopoiesis and lymphopoiesis. J. Exp. Med. *199*, 221–229.

Descamps, S., Arzouk, H., Bacou, F., Bernardi, H., Fedon, Y., Gay, S., Reyne, Y., Rossano, B., and Levin, J. (2008). Inhibition of myoblast differentiation by Sfrp1 and Sfrp2. Cell Tissue Res. *332*, 299–306.

Ferrer-Vaquer, A., Piliszek, A., Tian, G., Aho, R.J., Dufort, D., and Hadjantonakis, A.K. (2010). A sensitive and bright single-cell resolution live imaging reporter of Wnt/ß-catenin signaling in the mouse. BMC Dev. Biol. *10*, 121.

Gavard, J., Marthiens, V., Monnet, C., Lambert, M., and Mège, R.M. (2004). N-cadherin activation substitutes for the cell contact control in cell cycle arrest and myogenic differentiation: involvement of p120 and beta-catenin. J. Biol. Chem. *279*, 36795–36802.

Gould, S.J., and Lewontin, R.C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proc. R. Soc. Lond. B Biol. Sci. *205*, 581–598.

Günther, S., Kim, J., Kostin, S., Lepper, C., Fan, C.M., and Braun, T. (2013). Myf5-positive satellite cells contribute to Pax7-dependent long-term maintenance of adult muscle stem cells. Cell Stem Cell *13*, 590–601.

Han, X.H., Jin, Y.R., Seto, M., and Yoon, J.K. (2011). A WNT/beta-catenin signaling activator, R-spondin, plays positive regulatory roles during skeletal myogenesis. J. Biol. Chem. *286*, 10649–10659.

Harada, N., Tamai, Y., Ishikawa, T., Sauer, B., Takaku, K., Oshima, M., and Taketo, M.M. (1999). Intestinal polyposis in mice with a dominant stable mutation of the beta-catenin gene. EMBO J. *18*, 5931–5942.

Holland, J.D., Klaus, A., Garratt, A.N., and Birchmeier, W. (2013). Wnt signaling in stem and cancer stem cells. Curr. Opin. Cell Biol. *25*, 254–264.

Hutcheson, D.A., Zhao, J., Merrell, A., Haldar, M., and Kardon, G. (2009). Embryonic and fetal limb myogenic cells are derived from developmentally distinct progenitors and have different requirements for beta-catenin. Genes Dev. *23*, 997–1013.

Kim, C.H., Neiswender, H., Baik, E.J., Xiong, W.C., and Mei, L. (2008). Beta-catenin interacts with MyoD and regulates its transcription activity. Mol. Cell. Biol. *28*, 2941–2951.

Kramerova, I., Kudryashova, E., Wu, B., and Spencer, M.J. (2006). Regulation of the M-cadherin-beta-catenin complex by calpain 3 during terminal stages of myogenic differentiation. Mol. Cell. Biol. *26*, 8437–8447.

Kuang, S., Chargé, S.B., Seale, P., Huh, M., and Rudnicki, M.A. (2006). Distinct roles for Pax7 and Pax3 in adult regenerative myogenesis. J. Cell Biol. *172*, 103–113.

Kuroda, K., Kuang, S., Taketo, M.M., and Rudnicki, M.A. (2013). Canonical Wnt signaling induces BMP-4 to specify slow myofibrogenesis of fetal myoblasts. Skelet. Muscle *3*, 5.

Le Grand, F., Jones, A.E., Seale, V., Scimè, A., and Rudnicki, M.A. (2009). Wnt7a activates the planar cell polarity pathway to drive the symmetric expansion of satellite stem cells. Cell Stem Cell *4*, 535–547.

Lepper, C., Conway, S.J., and Fan, C.M. (2009). Adult satellite cells and embryonic muscle progenitors have distinct genetic requirements. Nature *460*, 627–631.

Lepper, C., Partridge, T.A., and Fan, C.M. (2011). An absolute requirement for Pax7-positive satellite cells in acute injury-induced skeletal muscle regeneration. Development *138*, 3639–3646.

Li, Y., Rankin, S.A., Sinner, D., Kenny, A.P., Krieg, P.A., and Zorn, A.M. (2008). Sfrp5 coordinates foregut specification and morphogenesis by antagonizing both canonical and noncanonical Wnt11 signaling. Genes Dev. *22*, 3050–3063.

Mauro, A. (1961). Satellite cell of skeletal muscle fibers. J. Biophys. Biochem. Cytol. *9*, 493–495.

Murphy, M., and Kardon, G. (2011). Origin of vertebrate limb muscle: the role of progenitor and myoblast populations. Curr. Top. Dev. Biol. *96*, 1–32.

Murphy, M.M., Lawson, J.A., Mathew, S.J., Hutcheson, D.A., and Kardon, G. (2011). Satellite cells, connective tissue fibroblasts and their interactions are crucial for muscle regeneration. Development *138*, 3625–3637.

Muzumdar, M.D., Tasic, B., Miyamichi, K., Li, L., and Luo, L. (2007). A global double-fluorescent Cre reporter mouse. Genesis *45*, 593–605.

Nastasi, T., Bongiovanni, A., Campos, Y., Mann, L., Toy, J.N., Bostrom, J., Rottier, R., Hahn, C., Conaway, J.W., Harris, A.J., and D'Azzo, A. (2004). Ozz-E3, a muscle-specific ubiquitin ligase, regulates beta-catenin degradation during myogenesis. Dev. Cell *6*, 269–282.

Niehrs, C. (2012). The complex world of WNT receptor signalling. Nat. Rev. Mol. Cell Biol. *13*, 767–779.

Otto, A., Schmidt, C., Luke, G., Allen, S., Valasek, P., Muntoni, F., Lawrence-Watt, D., and Patel, K. (2008). Canonical Wnt signalling induces satellite-cell proliferation during adult skeletal muscle regeneration. J. Cell Sci. *121*, 2939–2950.

Oustanina, S., Hause, G., and Braun, T. (2004). Pax7 directs postnatal renewal and propagation of myogenic satellite cells but not their specification. EMBO J. *23*, 3430–3439.

Pansters, N.A., van der Velden, J.L., Kelders, M.C., Laeremans, H., Schols, A.M., and Langen, R.C. (2011). Segregation of myoblast fusion and muscle-specific gene expression by distinct ligand-dependent inactivation of GSK-3β. Cell. Mol. Life Sci. *68*, 523–535.

Polesskaya, A., Seale, P., and Rudnicki, M.A. (2003). Wnt signaling induces the myogenic specification of resident CD45+ adult stem cells during muscle regeneration. Cell *113*, 841–852.

Relaix, F., Montarras, D., Zaffran, S., Gayraud-Morel, B., Rocancourt, D., Tajbakhsh, S., Mansouri, A., Cumano, A., and Buckingham, M. (2006). Pax3 and Pax7 have distinct and overlapping functions in adult muscle progenitor cells. J. Cell Biol. *172*, 91–102.

Ruiz-Herguido, C., Guiu, J., D'Altri, T., Inglés-Esteve, J., Dzierzak, E., Espinosa, L., and Bigas, A. (2012). Hematopoietic stem cell development requires transient Wnt/β-catenin activity. J. Exp. Med. *209*, 1457–1468.

Sambasivan, R., Yao, R., Kissenpfennig, A., Van Wittenberghe, L., Paldi, A., Gayraud-Morel, B., Guenou, H., Malissen, B., Tajbakhsh, S., and Galy, A. (2011). Pax7-expressing satellite cells are indispensable for adult skeletal muscle regeneration. Development *138*, 3647–3656.

Seale, P., Sabourin, L.A., Girgis-Gabardo, A., Mansouri, A., Gruss, P., and Rudnicki, M.A. (2000). Pax7 is required for the specification of myogenic satellite cells. Cell *102*, 777–786.

Shanely, R.A., Zwetsloot, K.A., Childs, T.E., Lees, S.J., Tsika, R.W., and Booth, F.W. (2009). IGF-I activates the mouse type IIb myosin heavy chain gene. Am. J. Physiol. Cell Physiol. *297*, C1019–C1027.

Tanaka, S., Terada, K., and Nohno, T. (2011). Canonical Wnt signaling is involved in switching from cell proliferation to myogenic differentiation of mouse myoblast cells. J. Mol. Signal. *6*, 12.

von Maltzahn, J., Chang, N.C., Bentzinger, C.F., and Rudnicki, M.A. (2012). Wnt signaling in myogenesis. Trends Cell Biol. *22*, 602–609.

von Maltzahn, J., Jones, A.E., Parks, R.J., and Rudnicki, M.A. (2013). Pax7 is critical for the normal function of satellite cells in adult skeletal muscle. Proc. Natl. Acad. Sci. USA *110*, 16474–16479.

Yi, L., and Rossi, F. (2011). Purification of progenitors from skeletal muscle. J. Vis. Exp. (49), pii: 2476.

Zhao, P., and Hoffman, E.P. (2004). Embryonic myogenesis pathways in muscle regeneration. Dev. Dyn. *229*, 380–392.

CHAPTER 4

HUMAN VARIANT PRIORITIZATION

VAAST variant prioritizer

Variant prioritization is the process of categorizing individual variants into groups based on some desired property. For example, often it is of research and medical interest to prioritize genetic variants according to how likely they are to contribute to disease. A major challenge of variant prioritization is that some genes naturally tolerate more variation than others, including missense and other protein coding variants. Thus, in order to successfully prioritize variants, the local genetic context of a variant is very important.

The NCBI's dbSNP database contains over 100 million human variants. Methods are needed that accurately and efficiently prioritize all known human genetic variants, not just those that induce a protein coding change or any other specified subset of variants; human genetic variants of nearly every conceivable annotation category have been associated with or shown to cause disease or phenotypic differences. Many software tools exist to prioritize human variants; however, they all suffer from significant limitations (Kircher et al., 2014). CADD is currently the most comprehensive tool available, and can prioritize SNVs and small insertion-deletion (indel) mutations (Kircher et al., 2014). However, CADD cannot process larger indels. To address the shortcomings of these other existing software tools, I have developed VVP, the VAAST Variant Prioritizer. VVP enables rapid, comprehensive, and accurate prioritization of all human variants. VVP is able to score all variation that can be annotated by Ensemble's Variant Effect Predictor (VEP) and, as I demonstrate below, is the fastest and most accurate tool available. VVP leverages the likelihood developed by Yandell et al. for VAAST and thus incorporates information about background allele frequency, amino acid change severity, and evolutionary conservation in order to prioritize human variation (Hu et al., 2013; Yandell et al., 2011). Because VVP incorporates allele frequency information in its scoring process, it is able to use zygosity information about the variants, which most other tools, including CADD, do not; thus, VVP is aware of dominant or recessive variation, which to my knowledge is not part of any other variant prioritization tool. VVP is implemented in Python and is available for academic use through the Yandell lab github repository.

VVP methodology

VAAST is a highly effective software tool that uses a burden test to identify genes responsible for disease (Rope et al., 2011). VAAST scores each genetic variant in the affected individuals using a likelihood equation that incorporates information about allele frequency in the target and background populations, amino acid change severity, and evolutionary conservation (Hu et al., 2013; Yandell et al., 2011). After scoring each variant using the likelihood, VAAST then filters through the scored variants to identify the highest scoring variant(s) that fit the specified penetrance and inheritance model (VAAST will choose one homozygous variant or two heterozygous variants for each target individual when a recessive model is specified). The VAAST gene burden score is then the sum of the scores of these identified variants. The statistical significance of the burden of a gene is determined by permuting the background and target populations. For full details on the VAAST methodology, see Yandell et al. (2011).

Although VAAST scores every variant using its likelihood, it does not provide a framework with which to prioritize individual variants. One cannot directly prioritize variants using the VAAST likelihood scores because there is no notion of the significance of the magnitude of the difference between any two scores. VVP overcomes this limitation by normalizing the VAAST likelihood scores into percentiles. This is done by calculating their percentile rank against three types of lookups that are built by cataloging healthy human variation in a background population. In this application, a lookup is defined as the percentile ranks of VAAST likelihood scores of healthy human variation. The three types of lookups are for coding variants, noncoding variants in a gene, and intergenic variants. Separate lookups for coding variants and noncoding variants are created for every gene and a single lookup is used for all intergenic variation. Through benchmarking, I have found this segmentation of the lookups works well, but it is a matter of further research to determine the best way to separate the lookups.

Suppose we have a genetic variant X in gene A with VAAST likelihood score of 9.2, and that gene A has a corresponding lookup Y. By comparing 9.2 to the percentiles of lookup Y, suppose we find that 9.2 has a percentile of 75. The VVP score of X is then 75. The interpretation of this result for variant X is that its score is greater than or equal to 75% of healthy human variants in gene A. In practice, a good cutoff is to consider variants with VVP scores higher than 98 to be potentially damaging (top 2% of variation). It is important to note that the lookups are entirely empirical, which means there are no parametric assumptions made about the shape or scale of the healthy human variation for any gene. I believe this is a strength of VVP, as there is very large variation in the shape of the distribution of scores in different genes (Figure 4.1).

The background human variation that is used to generate the lookups has a very large impact on the behavior and performance of VVP. Optimally, the background would have its variants called with the target variants of interest. However, I have used the 1000 genomes phase 3 variant calls as a general lookup with good success. Figure 4.2 shows that using variant calls from a background that was called with the target individuals to generate the lookups results in less noise in the VVP prioritization results. This is due to a higher relative VVP score in the background that was called with the target individuals than using the 1000 genomes phase 3 variant calls to generate the lookups. An important point brought out by Figure 4.2 is the comparability of VVP scores. The disease causing variant has a VVP score of 80 when using the lookups based on the 1000 genomes phase 3 variant calls and a VVP score of 100 when using the background that was called together with the target individuals; VVP scores are comparable to one another as long as the same background lookups are used to process the target variants of interest. However, VVP scores generated from different background lookups should not be compared since the VVP score is a lookup-specific measure of how extreme a variant score is.

Background lookup generation

In order to generate the background lookups to produce VVP scores, a vcf file of genotypes for the background individuals that has been annotated by VEP is required. Specific VEP annotation requirements are specified in the code distribution of VVP. The lookup generation then proceeds by scoring every individual with a variant genotype against every other individual in the vcf file using the VAAST likelihood. These scores are saved in separate bins for coding and noncoding variants for each gene. Intergenic scores are also saved in a separate bin. After processing all the variants for any particular feature, the lookup is created for each bin by calculating every percentile from 0 to 100 given the scores in the bin. These lookups are saved in an output file for use in scoring target variants.

Target variant scoring

Once background lookups have been generated, variants can be assigned VVP scores. The target variant file must also be in vcf format with VEP annotations. As in the background calculation, every variant genotype is scored using the VAAST likelihood. The percentile rank of the VAAST score is calculated using the appropriate background lookup. The current implementation of VVP will score target VCF files that have multiple individuals in them by scoring every individual genotype separately. A future direction is to combine VVP scores in the same gene from multiple individuals to calculate a burden score.

VVP results

Benchmarking was done with variants from the ClinVar database. I used ClinVar variants that were labeled as pathogenic or benign that had a known mode of inheritance of either dominant or recessive. Using this information, I was able to test VVP on variants that cause both recessive and dominant disorders and compare its results to both CADD and SIFT (Figure 4.3). Figure 4.3 shows VVP outperforms CADD or SIFT on this test dataset. VVP and CADD are able to score far more variants than SIFT (Figure 4.4).

CADD provides downloadable tables with precomputed scores for all SNVs and many small indels. However, as of writing this, the implementation of CADD is extremely slow and takes about a week to process the NA12878 vcf from 1000 genomes phase 3 data. However, given all the precomputed data, it is not difficult to imagine an implementation of CADD that scales well with growing datasets. SIFT scores can also be precomputed for all possible coding changes and thus also scales to large datasets (especially since SIFT scores a small fraction of possible human variation). VVP is also a very scalable approach since the background lookups need to be computed once and then target variants can be processed very quickly. VVP takes ~10 hours to process the entirety of NA12878 phase 3 vcf with 20 cpus. This time can be shortened further with the use of more processors. CADD's current implementation does not have the ability to utilize more cpus than its default operation, and therefore cannot take advantage of modern servers with many cpus.

VVP and CADD are currently the only variant prioritization tools with a broad ability to categorize human genetic variation. VVP has superior variant prioritization accuracy, can prioritize more indel and structural variation, and is much faster than the current implementation of CADD. Thus, VVP is currently the best single tool for broad human variant prioritization.

References

Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., & Yandell, M. (2013). VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, *37*(6), 622–34. doi:10.1002/gepi.21743

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–5. doi:10.1038/ng.2892

Rope, A. F., Wang, K., Evjenth, R., Xing, J., Johnston, J. J., Swensen, J. J., … Lyon, G. J. (2011). Using VAAST to identify an x-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency. *American Journal of Human Genetics*, *89*(1), 28–43. doi:10.1016/j.ajhg.2011.05.017

Yandell, M., Huff, C. D., Hu, H., Singleton, M., Moore, B., Xing, J., … Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, (21), 1529–1542. doi:10.1101/gr.123158.111

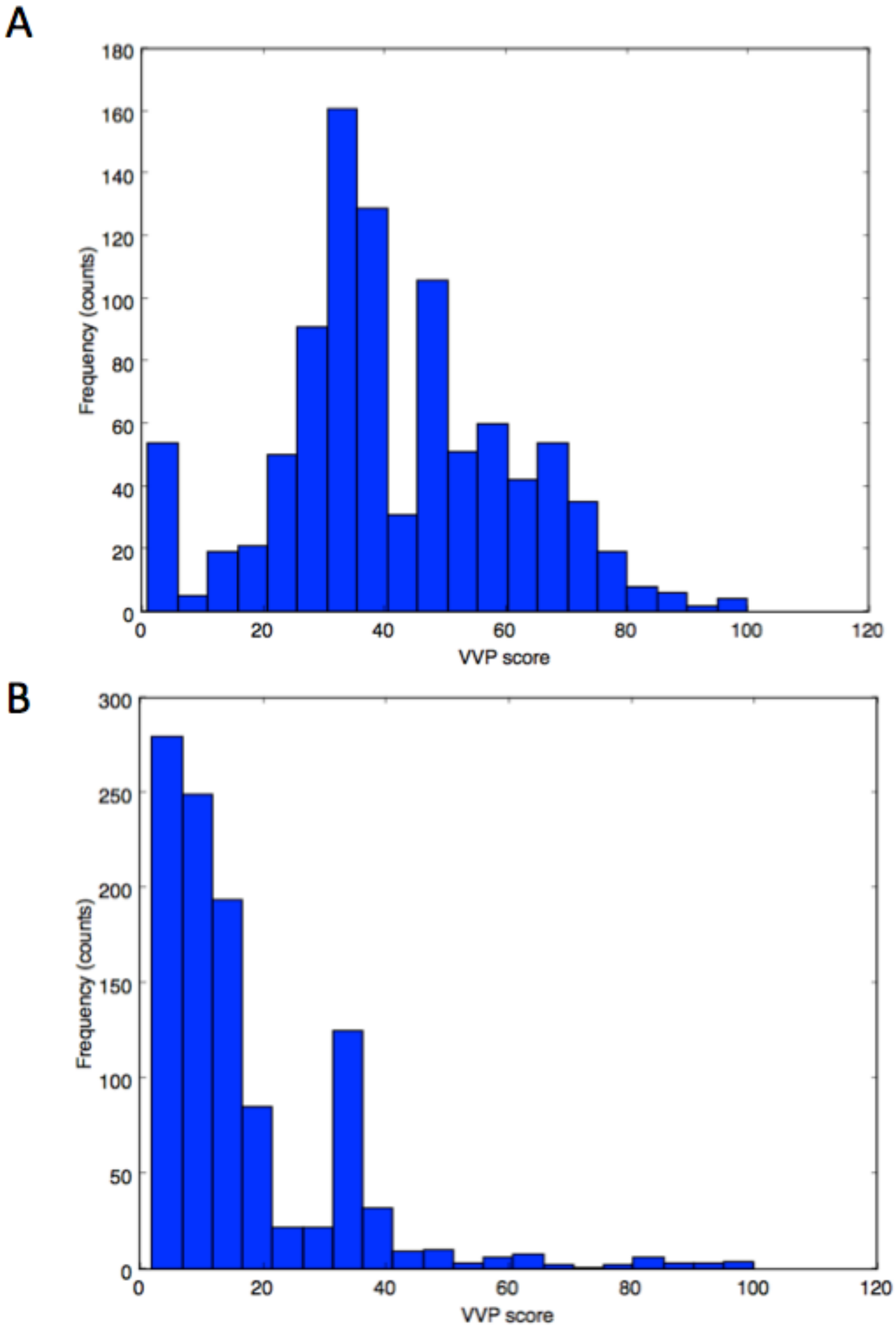


Figure 4.1: Histograms (A, B) are of different genes. Note the large difference in distribution of VVP scores between these genes. Most genes have very different distributions from one another. Not only are the distribution shapes highly variable, but also the relative number of variants in a gene. Some genes have thousands of known variants, while others may only have a handful. This is why, I believe, empirical lookups are better for prioritization than a parametric model.

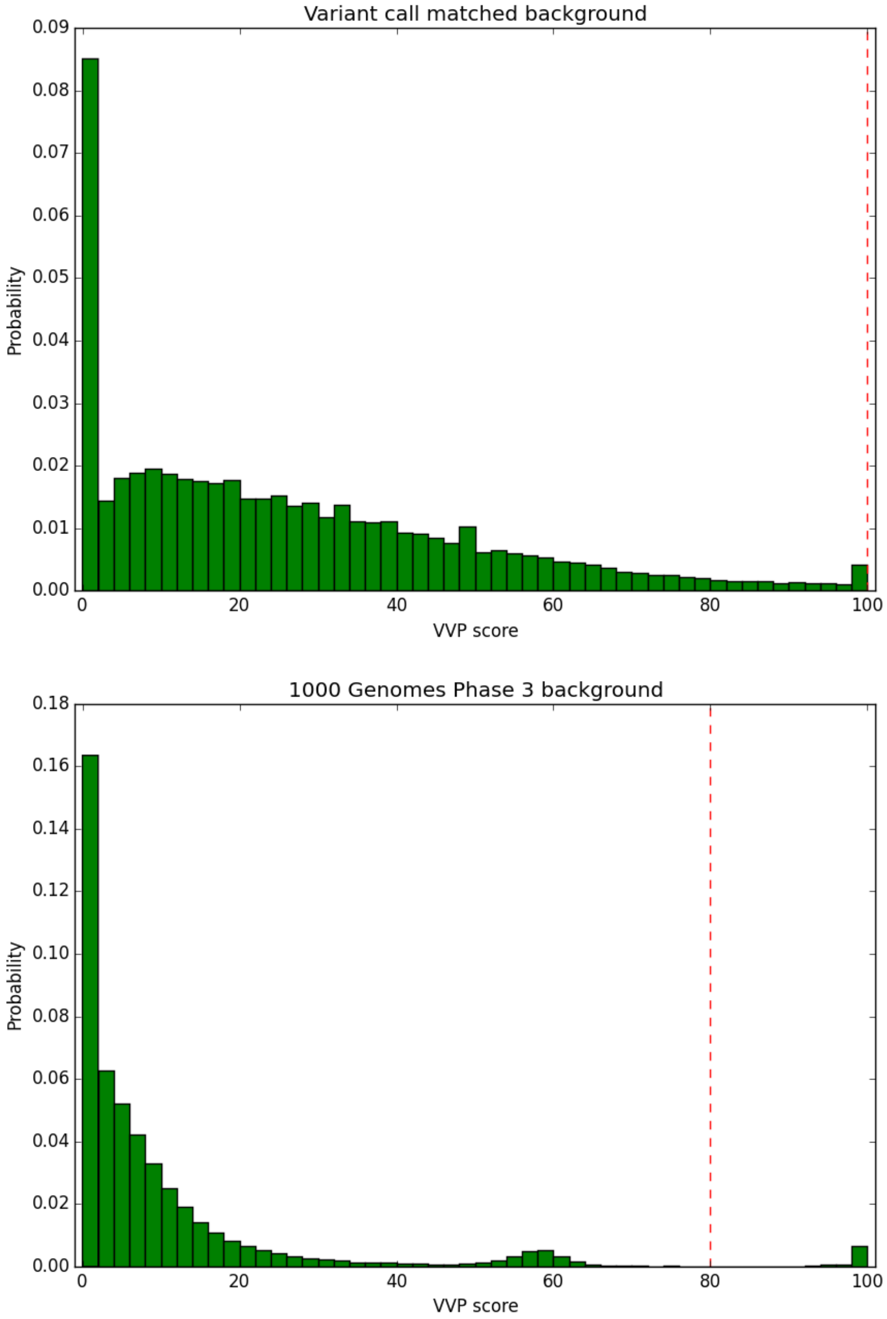


Figure 4.2: Histograms of VVP scores for 5 individuals that share a disease causing mutations in gene KCNQ1. Red dashed vertical line indicates score and relative position of known disease causing mutation. In the matched background (top panel), the signal is much stronger than in the background of the 1000 genomes phase 3 data (bottom panel).

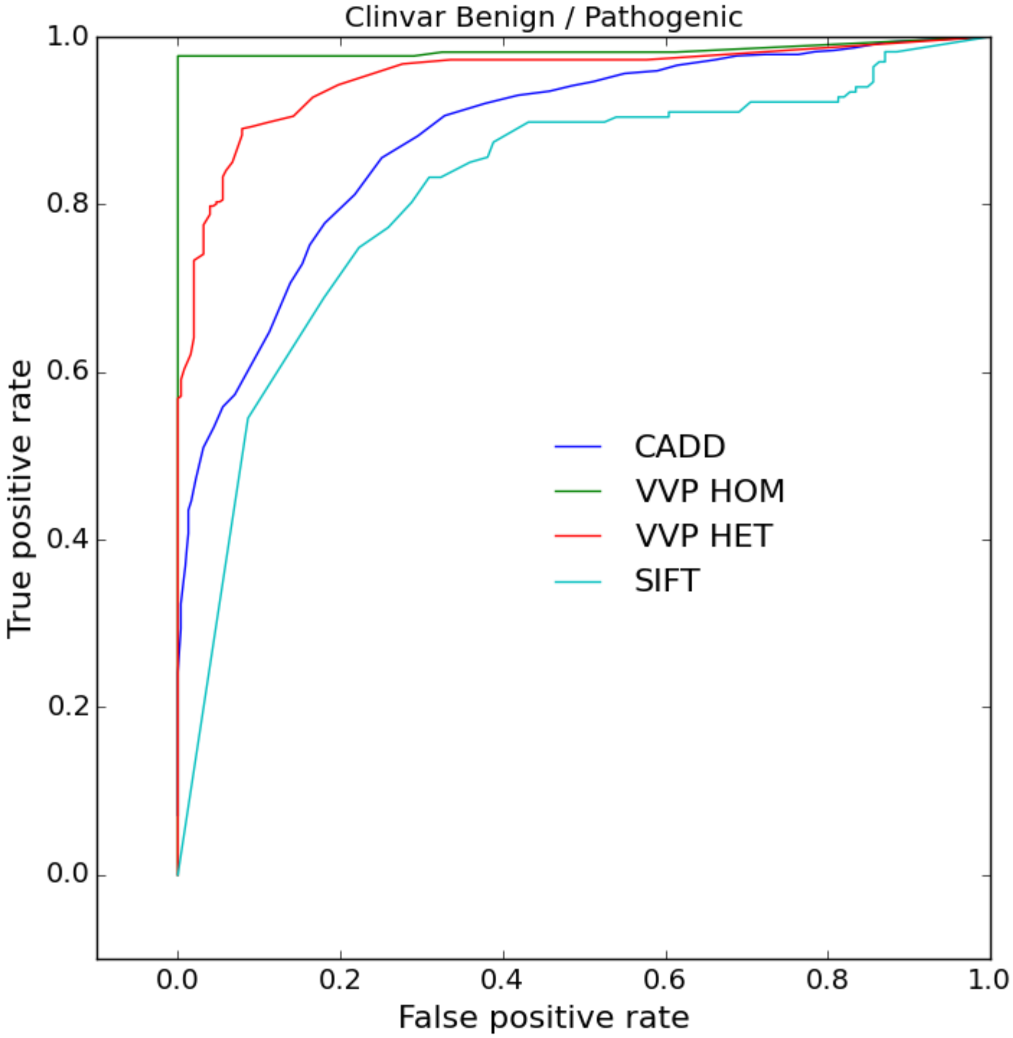


Figure 4.3: ROC curves for CADD, VVP, and SIFT. VVP is a better at discriminating between the pathogenic and benign ClinVar variants than CADD or SIFT. VVP is shown with its performance on homozygous and heterozygous variants since variants causing both recessive and dominant disorders are part of this benchmarking subset. Neither CADD nor SIFT distinguishes between homozygous and heterozygous variants so their performance is shown without dividing the variants by zygosity.

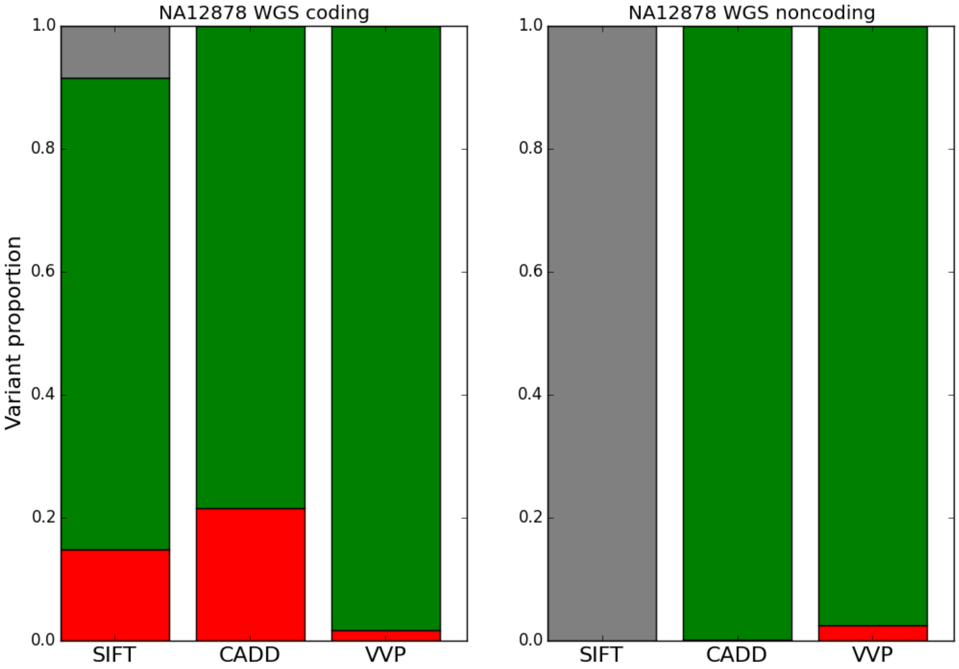


Figure 4.4: Stacked bar plots showing the classifications decisions on NA12878 variants. SIFT is only able to score a small subset of all variation. CADD predicts more variants to be damaging in the coding regions of this healthy individual.

CHAPTER 5

TAXONOMER: INTEGRATED MEANS FOR ULTRAFAST COMPREHENSIVE METAGENOMIC AND HOST TRANSCRIPTIONAL PROFILING

Steven Flygare, Keith Simmon, Chase Miller, Yi Qiao, Brett Kennedy, Tonya Di Sera, Erin H. Graf, Keith Tardif, Aurélie Kapusta, Chris Stockmann, Krista Queen, Suxiang Tong, Karl V. Voelkerding, Anne Blaschke, Carrie L. Byington, Seema Jain, Andrew Pavia, Krow Ampofo, Karen Eilbeck, Gabor Marth, Mark Yandell, and Robert Schlaberg

Contributions: I wrote all of the algorithmic code, helped in writing methods, was a main contributor to the extensive benchmarking, and generated much of the data presented in the results.

Introduction

Metagenomics, the genomic analysis of a population of microorganisms, makes possible the profiling of microbial communities in the environment and the human body at unprecedented depth and breadth. Its rapidly expanding use is revolutionizing our understanding of microbial diversity in natural and man-made environments and is linking microbial community profiles with health and disease (Afshinnekoo et al., 2015; Dickson, Martinez, & Huffnagle, 2014; Firth et al., 2014; Gilbert, Jansson, & Knight, 2014; Human Microbiome Project Consortium, 2012; Louis, Hold, & Flint, 2014; Mayer, Tillisch, & Gupta, 2015; Sherrard, Tunney, & Elborn, 2014; L. Zhao, 2013). To date, most studies have relied on PCR amplification of microbial marker genes (e.g., bacterial 16S rRNA), for which large, curated databases have been established ("The Greengenes Database. http://greengenes.secondgenome .com," n.d.; "UNITE," 2014; Yilmaz et al., 2014). More recently, higher throughput and lower cost sequencing technologies have enabled a shift towards enrichment-independent metagenomics. These approaches reduce bias, improve detection of less abundant taxa, and enable discovery of novel pathogens (Chiu, 2013; Lipkin, 2013; Shakya et al., 2013). In addition, they promise to revolutionize how infectious diseases are diagnosed and are of great interest for rapid, field-based biodefense testing. While conventional, pathogen-specific nucleic acid amplification tests are highly sensitive and specific, they require *a priori* knowledge of likely pathogens (i.e., they answer the question 'is pathogen X present'). The result is increasingly large, yet inherently limited diagnostic panels to enable diagnosis of the most common pathogens (Caliendo et al., 2013). Exhaustive follow-up testing may be required if first-line tests are negative. In contrast, enrichment-independent high-throughput sequencing allows for unbiased, hypothesis-free detection and molecular typing of a theoretically unlimited number of common and unusual pathogens (i.e., answering the question 'what pathogen is present'). Unbiased, sequencing-based pathogen detection has led to the diagnosis of previously unrecognized infections and discovery of novel pathogens in select cases (see Wilson et al., 2014 for example). Its wide adoption is likely to revolutionize the laboratory diagnosis of infectious diseases and will aid in the rapid response to public health emergencies.

While direct pathogen identification from high-throughput sequencing data is generally the goal, other analysis modalities are possible. Differentiating viral from bacterial infections, for example, can indicate whether antibiotic treatment is necessary. This has traditionally been attempted through phenotyping of the host leukocyte response (e.g., leukocyte count, differential cell count) or protein markers (e.g., C-reactive protein, procalcitonin). More recently, microarray-based host transcript expression profiling from blood leukocytes has been used to demonstrate proof-of-concept for differentiating infectious etiologies (X. Hu, Yu, Crosby, & Storch, 2013a; Zaas et al., 2013, 2009). Here too, high-throughput sequencing has much to offer. The greater sensitivity and unbiased nature of RNA-seq enables simultaneous pathogen detection and host-response profiling. Such data could be used to better inform treatment, potentially overcoming many of the limitations of current infectious disease tests (Caliendo et al., 2013; Hudson, Woods, & Ginsburg, 2014).

Wide availability of next-generation sequencing instruments, lower reagent costs, and streamlined sample preparation protocols have enabled an increasing number of investigators to perform high-throughput DNA and RNA-seq for metagenomics studies. However, analysis of sequencing data is still forbiddingly difficult and time consuming, requiring bioinformatics skills, computational resources, and microbiological expertise that is not available in many laboratories, especially diagnostic ones. Clearly, more computationally efficient, accurate, and easy-to-use tools for comprehensive diagnostic and metagenomics analyses are needed.

Here we describe Taxonomer, an integrated, ultrafast tool for metagenomic sequence analysis. Taxonomer enables novel analysis modalities of unmatched complexity in an easy-to-use format, including the following: (1) comprehensive panmicrobial detection and discovery, (2) host-response profiling, (3) interactive result visualization, and (4) access through a web-based user interface, which eliminates the need for specialized hardware or expertise. Taxonomer operates at speeds comparable to the fastest, ultrafast tool Kraken (up to 4 million reads per minute), but unlike Kraken, Taxonomer supports both nucleotide and protein-based classification using a single integrated algorithmic framework (Wood & Salzberg, 2014). This means that Taxonomer can be used for many additional applications such as virus detection and phylogenetic classification, while providing greater accuracy and comprehensive taxonomic profiling at 1-2 orders of magnitude faster classification speeds than alignment-based tools such as those used by SURPI (Naccache et al., 2014). Moreover, Taxonomer also enables new analysis modalities that are crucial for understanding both complex metagenomic data and for developing unbiased diagnostic approaches. Taxonomer can be used in the analysis of DNA and/or RNA (total or poly-A selected) sequencing; it is not restricted to short reads (i.e., can be used to analyze contigs assembled from metagenomics datasets); and is the only ultrafast metagenomics tool that provides integrated means for quantification of human transcripts, allowing simultaneous identification of pathogens, assessment of their relative abundance, and quantification of the patient's transcriptional response to the infection.

Taxonomer is the result of a multidisciplinary effort and enables these applications through a set of four integrated tools (Binner, Classifier, Protonomer, and Afterburner) (Figure 5.1a; see methods for details). Collectively, these four interlocking modules provide synergistic means for nucleotide and protein-based homology searches, phylogenetic classification, and host transcriptional profiling. Taxonomer is available via an iobio web-service (Figure 5.1b), allowing rapid, highly interactive analyses accessible through personal computers and mobile devices without the need for special computational infrastructure on the user side (Miller, Qiao, DiSera, D'Astous, & Marth, 2014).

Here we demonstrate the power of Taxonomer using both, synthetic and biological data sets, and evaluate its speed and classification accuracy by comparing it to state-of-the-art tools for sequence alignment (BLAST), rapid metagenomic data analysis (Kraken, SURPI), marker gene-based microbial classification (RDP Classifier), protein searches (RapSearch2, DIAMOND), and RNA-seq-based transcriptional profiling (Sailfish, and Cufflinks) (Altschul, Gish, & Miller, 1990; Buchfink, Xie, & Huson, 2015; Cole et al., 2014; Naccache et al., 2014; Patro, Mount, & Kingsford, 2014; Trapnell et al., 2010; Wood & Salzberg, 2014; Y. Zhao, Tang, & Ye, 2012). As we demonstrate, Taxonomer is ultrafast, more accurate, and more comprehensive in scope, and enables new modalities of analysis for clinical metagenomics datasets not provided by any other tool.

Methods

Binner module

Identifying small numbers of pathogen sequences hidden among vast numbers of host and/or microbiota-derived sequencing reads is a major algorithmic challenge for metagenomics-based pathogen detection tools. The standard approach is to use digital subtraction (Borozan, Watt, & Ferretti, 2013), whereby all sequencing reads are first aligned to the host's genome sequence. This is the approach used by SURPI (Naccache et al., 2014), for example. During subtraction, reads of host origin are removed. Additional subtraction steps may be used for removal of nonrelevant microbial sequences, including those known to represent reagent contamination or sequencing adaptors (Gire et al., 2014). A greatly reduced number of presumably relevant microbial sequences are then classified by alignment to larger reference databases. Since only the remaining reads are matched with selected reference sequences, pathogens can be missed entirely if they are homologous to sequences in the subtraction database. Taxonomer overcomes this inherent limitation of digital subtraction by means of its 'Binner' module (Figure 5.1a), which compares each read to every reference database in parallel, assigning them to broad, nonexclusive taxonomic categories.

Taxonomer's binner database is created by counting unique 21bp k-mers in different taxonomic/gene datasets using Kanalyze (version 0.9.7) (Audano & Vannberg, 2014). Each taxonomic/gene dataset represents a 'bin' in which query sequences can be placed based on their k-mer content. Each database is assigned a unique bit flag that allows k-mers that belong to one or more bins to be recognized and counted. The k-mer counts are merged into a binary file that contains the k-mers and the database flag. This binary file shares a similar organization to our classification databases, and is organized to optimize query speed. Reads are then assigned to the taxonomic group(s) with which most k-mers are shared. Ties are resolved based on the bins we expect the majority of sequences to arise from. High binning accuracy is possible because of the minimal intersections (0.47%) of k-mer content from comprehensive human and microbial reference databases. Optimal k-mer count cutoffs were determined by Youden's indexes and F1 scores and ranged from 3 to 13 (Akobeng, 2007). To eliminate binning of reads containing adapter sequence, by default, the binner ignores k-mers present in Illumina Tru-Seq adapters. A database of External RNA Controls Consortium (ERCC) control sequences allows quantification of ERCC spike-in controls.

Classifier module

Classification in Taxonomer is based on exact k-mer matching. Taxonomer uses databases that are optimized for rapid k-mer queries that store every reference in which a k-mer is found as well as an associated k-mer weight for every reference. The fundamental question for classification is how likely it is that a particular k-mer (K_i) originates from any reference sequence, ref_i. To answer this question, Taxonomer calculates a k-mer weight:

$$KWref_i(K_i) = \frac{C_{ref}(K_i)/C_{db}(K_i)}{C_{db}(K_i)/Total\ kmer\ count}$$

where C represents a function that returns the count of K_i. $C_{ref}(K_i)$ indicates the count of the K_i in a particular reference. $C_{db}(K_i)$ indicates the count of K_i in the database. This weight provides a relative, database specific measure of how likely it is that a k-mer originated from a particular reference. In order to classify a query sequence, we calculate the sum of the k-mer weights for every reference that has a matching k-mer in the query sequence. Suppose that there are N possible k-mers from query sequence Q. Then, for every reference, ref_i, that shares a k-mer with Q, the total k-mer weight for ref_i is:

$$TKW(ref_i) = \sum_{j=1}^{N} KWref_i\,(K_j)$$

Each read is assigned to the reference that has the maximum total k-mer weight. In the case of a tie, the query sequence is assigned to the taxonomic lowest common ancestor (LCA).

Protonomer module

We developed a mapping scheme between amino acids and their corresponding codons to facilitate mapping in protein space while using the same strategies and speed we developed for classification in nucleotide space. When the amino acid database is built for classification, Taxonomer assigns every amino acid to just one codon. This unique mapping, which we term a *non-degenerate translation,* is used to generate an artificial DNA sequence that corresponds to the protein sequence in the database. This DNA sequence is entered into Taxonomer's nucleotide classification databases. Query reads are translated into all 6 reading frames using the same non-degenerate translation scheme used to build the database and each translated frame is then classified. K-mer weighting and read classification assignment are performed as described above. The default Protonomer database is a subset of UniRef90 (see Databases for details). Empirically, we found a k-mer size of 30 (10 amino acids) to perform best. We chose to classify viruses in protein space because of their high mutation rates, genetic variability, and incomplete reference databases (Anthony et al., 2013). Figure 5.2 presents benchmark data for Protonomer and two other rapid protein search tools, RAPSearch2 (employed by SURPI) and DIAMOND (an ultrafast, BLAST-like protein search tool), using RNA-seq data from respiratory samples of 24 children with documented viral infections as determined by an FDA-cleared molecular test (eSensor Respiratory Virus Panel, GenMark) for which complete viral genomes could be manually constructed (Buchfink et al., 2015; Y. Zhao et al., 2012). Viral reads were defined by mapping all reads binned as 'Viral' or 'Unknown' to the manually constructed viral genomes (Geneious, version 6.1). Sensitivity and specificity were determined based on detection of known viral reads (true positives) and nonviral reads (true negatives). Protonomer provides a single taxonomic identifier per read as the classification assignment, which makes interpretation of results extremely simple. Neither RAPSearch2 nor DIAMOND classify a read; instead, they only provide blast-like alignment information. For benchmarking against RAPSearch2 and DIAMOND, the LCA of the alignment with the lowest E-value was assigned as the classification. All tools were benchmarked using the same (Taxonomer's default) reference sequences as their database. Both Protonomer and RAPSearch2 process paired reads by concatenating them together with a '-

' between mate pairs. DIAMOND does not support paired end reads, so each pair was searched separately, and the hit with the lowest e-value from each read was used to make the classification assignments.

Afterburner

To increase recovery of distantly homologous viral proteins, Taxonomer offers two options. First, unclassified reads can be further analyzed using the Afterburner module, a degenerate k-mer matching engine that employs a collapsed amino-acid alphabet. In a manner similar to that employed by DIAMOND, we used k-means clustering on the BLOSUM62 matrix to generate a compressed amino acid alphabet (Buchfink et al., 2015). By using the collapsed amino acid alphabet, we are able to achieve higher sensitivity in classification with sequences that are more diverged at the expense of a higher false positive rate when compared with Protonomer. In addition, the Taxonomer package provides utility scripts to manufacture relevant read subsets for *de novo* assembly. Importantly, Taxonomer is not restricted to short reads, allowing re-analysis of resulting contigs for still greater classification sensitivity (Figure 5.2).

Host gene expression estimations

Taxonomer also uses its nucleotide classifier to assign reads to host reference transcripts. By default, these are transcripts and corresponding gene models (GTF file) from the ENSMBL human reference sequence, GRCh37.75. Empirically, we found that a k-mer size of 25 worked best for mapping reads to human transcripts. We benchmarked Taxonomer's gene expression estimates against Sailfish's and Cufflinks' using both biological and synthetic data (Patro et al., 2014; Trapnell et al., 2010). To generate the benchmark data shown in Figure 5.3a, we ran Taxonomer in a standalone fashion. We had Taxonomer output all ties between transcripts during the classification step; we then randomly assigned a read to a single transcript. We used these transcript level assignments to calculate gene level expression. We next employed a linear regression to correct for transcript assignment bias in a similar fashion to Sailfish. The reported correlations were then calculated using these corrected values. This level of gene expression analysis is not currently available through the web interface because of the way data are streamed; however, the results given from the web interface are a very good approximation (Spearman correlation > 0.93 on a set of genes that both methods have positives counts and Spearman correlation > 0.75 when the gene set is unrestricted). In the first experiment, we employed qPCR results taken from the microarray quality control study (MAQC)[38]; specifically, human brain tissue samples (Accession numbers SRR037452, SRR037453 , SRR037455 , SRR037455 , SRR037458). We also compared performance using synthetic RNA-seq reads (2x76bp, n=15,000,000) generated with the Flux Simulator tool. TopHat was used to produce alignments for Cufflinks (Griebel et al., 2012; Trapnell, Pachter, & Salzberg, 2009). Like Taxonomer, Sailfish does not need external alignment information.

Databases

The Classifier and Protonomer databases are modular and easily constructed, consisting only of multi-fasta files with a ‘parent tag’ on their definition lines. These tags describe each reference sequence’s immediate phylogenetic parent-taxon. **Bacterial classification** is based on a marker gene approach (16S rRNA gene) and the Greengenes database (reference set with operational taxonomic units, OTU, clustered at 99%, version 13_8 (DeSantis et al., 2006; McDonald et al., 2012). This reference set contains 203,452 OTU clusters from 1,262,986 reference sequences. The taxonomic lineage for each OTU was used to create a hierarchical taxonomy map to represent OTU relationships. To support the OTU ‘species’ concept, the taxonomy was completed for ranks in the taxonomic lineage that had no value. Unique dummy species names from the highest taxonomic rank available were used to fill empty values. Versions of the Greengenes database were formatted for use within BLAST, the RDP Classifier, and Kraken. **Fungal classification** is also based on a marker gene approach (internal transcribed spacer, ITS, rRNA sequences) and the UNITE database (version sh_taxonomy_qiime_ver6_dynamic_s_09.02.2014) (Koljalg et al., 2013). This reference set contains 45,674 taxa (species hypothesis, SH) generated from 376,803 reference sequences with a default-clustering threshold of 98.5% and expert taxonomic curation. Dummy names were created for ranks that had no value. Versions of the unite database were formatted for use with BLAST, the RDP Classifier, and Kraken. **Viral classification and discovery** is done using the protein sequences from UniRef90 downloaded on June 16, 2014. The database was reduced to 289,486 viral sequences based on NCBI taxonomy. Phage sequences were separated, leaving a total of 200,880 references for other viruses. NCBI taxonomy was used to determine the sequence relationship. For testing purposes, additional bacterial classification databases were constructed from RefSeq (identical to Kraken's full database; *n*=210,627 total references; *n*=5,242 bacterial references, using NCBI taxonomy), and the complete ribosomal database project databases download on September 24, 2014 (*n*=2,929,433 references, using RDP taxonomy).

Database construction

Databases are constructed to maximize query speed. K-mers are stored in lexicographical order and k-mer minimizers are used to point to blocks of k-mers in the database. Once a block of k-mers is isolated, a binary search is used to complete the query. This scheme provides extraordinary query speeds, as demonstrated by (Wood & Salzberg, 2014). We employ the same basic database layout as Kraken, with the important difference that instead of storing just the LCA of a k-mer, we also store the k-mer count and every reference (up to an adjustable cutoff) with associated k-mer weight. Detailed information about the database format and layout is available upon request.

Gene classification protocols

We extracted reference sequences from widely used, curated public databases for benchmark experiments (Yilmaz et al., 2014). These reference sequences were used to generate synthetic read datasets having a variety of read-lengths and error rates using wgsim. PCR-amplified 16S rRNA gene sequences from two metagenomics studies on stool and the home environment were also used (Lax et al., 2014; Subramanian et al., 2014). The analysis was limited to taxa with relative abundance >0.1% per sample (10 random samples were selected from each study).

Bacterial 16S rRNA

From the SILVA 119 nonredundant small-subunit ribosomal sequence reference database, we extracted bacterial reference sequences between 1200-1650bp of length and excluded references annotated as cyanobacteria, mitochondria, and chloroplasts (Yilmaz et al., 2014). Only high-quality references without ambiguous bases, alignment quality values >50%, and sequence quality >70% were included. All the above values are reported by SILVA. Percent identity to the closest Greengenes OTU was determined by MegaBLAST using hits with a query coverage >80% (Zhang, Schwartz, Wagner, & Miller, 2000). Synthetic reads (100bp single-end, 100bp paired-end, 250 paired-end) were generated from these reference sequences at 5X coverage.

Fungal ITS

To test the accuracy of identifying fungal ITS sequences that are not represented in the UNITE database, we utilized the UNITE_public_dataset (version_15.01.14) (Koljalg et al., 2013). Percent identity to the closest UNITE species hypothesis (SH, OTU's clustered at 98.5%) was determined by MegaBLAST using hits with a query coverage >80%. Synthetic reads (250bp single-end) were generated from these reference sequences at 5X coverage. Due to the variable length of ITS sequences (mean 585bp, range 51-2,995bp, *n*=376,803), paired-end sequences were not generated.

Classification criteria for reference methods

BLAST

Default MegaBLAST parameters were used. Top scoring references were identified and used to assign OTUs/SHs. Multiple OTUs/SHs were assigned to synthetic reads when more than one OTU/SH reference shared 100% identity. If no OTU/SH had 100% identity to a read, then all

OTUs within 0.5% of the top hit were assigned to the read. The taxonomy of the assigned OTUs/SHs was compared and the highest rank in common was used to assign a taxonomic value to the read. The percent identity was used to determine the assignment of the highest taxonomic rank. Sequence reads with >97% identity to a reference were assigned to species, >90% identity to genus, and <90% to family when lineage information was available at this rank.

RDP Classifier

RDP Classifier analyses were performed on a local server (see below). Classifications were resolved to the rank with a minimum confidence level of ≥0.5.

Kraken

Kraken analyses were performed on a local server (see below). Kraken reports the taxon identifier for each read's final taxonomic assignment.

SURPI

SURPI analyses were performed using an Amazon EC2 instance through the published Amazon Machine Image. SURPI reports the best hit for its mapping tools (SNAP, RAPSearch2), which were used for comparison (Zaharia et al., 2011).

Taxonomer implementation

Taxonomer was written in C with Python bindings through Cython. An implementation of Taxonomer that contains the entire pipeline functionality was written in C and drives the iobio web interface.

Server specifications

Benchmarking was performed on a machine with Red Hat Linux, 1TB of RAM, and 80 CPUs. Number of CPUs was restricted to 16 unless otherwise noted.

Web-service and visualization

Taxonomer is publically available as a web-service built upon the iobio framework (Miller et al., 2014). It is available at taxonomer.iobio.io. Complex metagenomic data can be processed quickly and effectively interpreted through web-based visualizations. Figure 5.1b illustrates the interface. As reads are being streamed to the analysis server, a pie chart is presented summarizing the results of the binning procedure. When one of the bacterial, fungal, viral, or phage bins of the pie chart is selected, the results of the Classifier/Protonomer modules are displayed in a sunburst visualization. Additional information is provided at the top of the web page about how many reads were sampled, the number of reads classified, and the detection threshold. The detection threshold informs a user about how abundant a particular organism must be in order to be detected with the number of reads sampled. This provides an indicator of the sensitivity of detection in the sample. In addition, a slider allows the user to select an absolute cutoff for the minimum number of reads required in order to be displayed in the sunburst.

DNA and RNA-seq of patient samples

Nucleic acid extraction

Samples (75-200µL) were extracted using the QIAamp Viral RNA extraction kit (Qiagen). Extraction was carried out as described by the manufacturer with the exception of the AW1 washing step. For this step, 250µL of AW1 wash buffer was added to the QIAamp Mini column before centrifugation at 8000 rpm. Then, 80µL of DNase I mix (Qiagen) containing 10µL of RNase-free DNase I and 70µL of Buffer RDD was added to the column for on column DNase digestion. After incubation at room temperature for 15 min, an additional 250µL of AW1 was added to the column before centrifugation at 8000 rpm. The manufacturer suggested protocol was continued at this point with column washing using Buffer AW2. After all washing steps, RNA was eluted in 60µL of water. Extraction for total DNA was performed using 75-200µL of sample with the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer’s instructions. DNA was eluted in 200 µL of nuclease-free water.

Depletion of human DNA

Microbial DNA was enriched with NEBNext Microbiome DNA Enrichment Kit (NEB). Briefly, MBD2-Fc-bound magnetic beads were prepared by combining 3µL of MBD2-Fc protein with 30µL of Protein A Magnetic Beads per sample and placing the mixture in a rotating mixer for 10 min at room temperature before washing with 1X Binding Buffer. Extracted DNA (200ng in 200µL) was added to 50µL 5X Binding Buffer. The resulting 250uL were added to MBD2-Fc-bound magnetic beads for 15 min at room temperature with rotation. The enriched microbial DNA was cleaned-up with Agencourt AMPure XP Beads (Beckman Coulter).

Library generation

For HiSeq and MiSeq sequencing, indexed cDNA libraries were produced from extracted RNA using the TruSeq RNA Sample Prep Kit v2 (Illumina) omitting poly-A selection. RNA was dried and resuspended in 19.5 µL of Elute, Prime, Fragment Mix. The remainder of the library preparation was conducted per manufacturer's instructions. Before library generation from DNA, enriched microbial DNA was fragmented with the Covaris S2 Ultrasonicator using intensity 5, duty cycle 10%, and 200 cycles/burst for 80 seconds all at 7 °C. Libraries generated from fragmented enriched microbial DNA were prepared using the KAPA Hyper Prep Kit (KAPA Biosystems) according to the manufacturer's instructions. PCR cycles used for library amplification were dependent upon the amount of input DNA and 13 cycles were used for these experiments. Libraries were quantitated by qPCR using the KAPA SYBR FAST ABI Prism qPCR Kit (KAPA BioSciences) and the Applied Biosystems 7900HT Fast Real-Time PCR System (Applied Biosciences). Library size was determined with the Agilent High Sensitivity DNA Kit and Agilent 2100 Bioanalyzer. After pooling of indexed sequencing libraries, a second qPCR and bioanalyzer run was performed to estimate the final concentration before sequencing. For Ion Proton sequencing, indexed cDNA libraries were produced from extracted RNA using the SMARTer Universal Low Input RNA Kit (Clontech) with numbers of PCR cycles ranging from 10-15 based on RNA yield.

Sequencing

Pooled sequencing libraries were analyzed on a HiSeq 2500 (2x100bp), MiSeq (2x250bp, both Illumina), or Ion Proton (median read length 139bp, Life Technologies) instruments according to manufacturers' protocols.

Statistical analyses

For gene expression analyses, we report both the Pearson and Spearman correlations as was done before (Patro et al., 2014). The Pearson correlation of the log transformed gene expression estimates necessitates the removal of any genes whose estimated expression is 0. The log transform prevents outliers from dominating the correlation. We also report the Spearman correlation, for which the log transform is not as necessary since it is a correlation based on ranks. Thus, the exclusion of genes with estimates of 0 can be avoided.

Results

Below, we present a series of benchmark analyses using biological and synthetic datasets; these include a large number of pediatric respiratory samples from the Centers for Disease Control and Prevention (CDC) Etiology of Pneumonia In the Community (EPIC) study as well as published data (Gire et al., 2014; Grard et al., 2012; Y. Hu et al., 2013; Jain et al., 2015). Our benchmark comparisons to other ultrafast tools for metagenomic classification, such as Kraken and SURPI as well as more established analysis tools, such as BLAST and RDP Classifier, demonstrate Taxonomer's speed and accuracy, and how it enables new analysis modalities.

Non-greedy binning

To demonstrate the advantage of Taxonomer's non-greedy binning algorithm, we compared high-level taxonomic assignments made by SURPI (which employs a greedy digital subtraction approach using SNAP) to those of Taxonomer's Binner for RNA-seq data (Zaharia et al., 2011). While high-level taxonomic assignments agree for 73.8% of reads, Taxonomer assigned 16% of reads an ambiguous origin (i.e., they match equally to multiple databases), 96% of these were classified as human by SURPI. This was mostly due to highly conserved ribosomal and mitochondrial sequences (data not shown), but similar effects were also apparent for fungal sequences (18% classified as human by SURPI). Taxonomer's alignment-free binning approach was also able to capture more phage/viral sequences (7,426) than the alignment-based method (5,798), and resulted in fewer unclassified sequencing reads (3.2% vs. 4.5%). Consistent with lower abundance of rRNA and mtRNA sequences in DNA sequencing data, Taxonomer had many fewer ambiguous assignments (0.04%, of which 40% were classified as human and 59% as viral by SURPI; overall agreement 98.7%). In addition to decreased numbers of false negatives, the Binner also provides users of the Taxonomer web-service with a high-level overview of the contents of even the largest and most complicated dataset within the first second or so of computation.

Analysis time and completeness of classification

Table 5.1 presents time and classification percentages for Taxonomer, Kraken, and SURPI. For this analysis, we used RNA-seq data from three virus-positive respiratory tract samples with a range of host vs. microbial composition profiles (Graf, 2015). Kraken was the fastest tool requiring about 1.5 min/sample on average, but because it relies on nucleic acid-level classification only and uses a single reference database, it classified fewer reads than Taxonomer and SURPI. Although SURPI enables amino acid-level searches for virus detection and discovery, this greatly extended analysis times to between 1.5 and >12 hours. Like SURPI, Taxonomer provides both nucleic acid and protein-based microbial classification. Taxonomer also automatically creates host gene expression profiles. Moreover, all these analyses are carried out very quickly; Taxonomer achieved times similar to Kraken requiring on average ~5 minutes to classify 5-8x10^6 paired-end reads using 16 CPUs. Moreover Taxonomer classified the largest number of reads in 2 of the 3 samples and tied with SURPI for the third sample. Collectively, these results provide an introduction and overview of how Taxonomer combines the ultrafast speed of Kraken with an extended suite of analysis and search capabilities that exceed those of SURPI.

Bacterial and fungal classification accuracy

A comprehensive classification database is essential for mitigating errors resulting from imperfect matches to query sequences. RefSeq is one solution, but it contains only some 5,000 sequenced bacterial taxa (at the time of access), whereas available 16S rRNA sequences suggest existence of at least 100,000 to 200,000 OTUs given existing sequence databases (Cole et al., 2014; McDonald et al., 2012; Yilmaz et al., 2014). Reads derived from taxa that are absent from the classification database can result in false negative and false positive classifications, especially at the genus and species level. Performance of classification tools is frequently only tested with synthetic reads derived from the reference database; i.e., perfect matches exist for all synthetic reads. For microbial classification, this is a highly artificial challenge, as novel species or strains are routinely encountered in clinical or environmental samples.

To provide a more realistic challenge, we generated synthetic reads from bacterial 16S rRNA sequences in the SILVA database lacking perfect matches in Taxonomer's Greengenes-derived reference database (468 of 1013 source references, 46%, had no perfect match in the classification database) (Yilmaz et al., 2014). This is why Taxonomer employs a marker gene approach and a custom Greengenes-derived database for prokaryotic classification.

The utility of Taxonomer's approach is illustrated in Figure 5.4a, demonstrating that SURPI, Kraken, and Taxonomer differ greatly as regards accuracy when using their default databases and command lines to classify error-free, synthetic 16S rRNA-derived reads. At the species level, for example, Taxonomer correctly classifies 59.5%, incorrectly classifies 15.7%, and fails to classify 24.8% of the reads. By comparison, Kraken classifies 29% of the reads to the correct species, and exhibits a high false positive rate, classifying every remaining read (71%) incorrectly. The results for SURPI have been split into two columns reflecting the fact that SURPI, unlike Taxonomer and Kraken, classifies each read from a mate pair independently, and in many cases, these assignments are discordant. Thus, the right-hand portion of the SURPI column records the classification rates when either read from a mate pair is classified correctly; the left-hand portion records the rates for classifying both mates to the same taxon. As can be seen, SURPI underperforms both Taxonomer and Kraken.

Figure 5.4b shows performance comparison of Taxonomer with the RefSeq (Kraken default), RDP, and Greengenes (Taxonomer default) databases. Using its default database, Taxonomer correctly classifies 59.5% of the reads, and recovers 94.9% of species. Using Kraken's default database (RefSeq DB), Taxonomer's values drop to 27% and 71.6%, respectively, similar to Kraken's results when using the same database: 29% and 71%, respectively. Also presented in Figure 5.4b are Taxonomer's classification and recovery rates using the RDP database (Cole et al., 2014). Although Taxonomer misclassified very few reads using the RDP database, overall performance was substantially better using Taxonomer's default database.

Figure 5.4c shows benchmarks for four different classification tools, MegaBLAST, the RDP Classifier, Kraken, and Taxonomer, all using Taxonomer's default 16S database (Cole et al., 2014; Sayers et al., 2010). SURPI is not included in this panel, as it provides no means for employing user-provided databases. Overall, Taxonomer's performance closely approximates that of the RDP Classifier, an established reference tool. At the species level, Taxonomer and RDP classify 59.5% and 61.4% of reads correctly, and recovery rates are very similar. Note that Kraken's classification and recovery rates improve dramatically using Taxonomer's database compared to its own, but that Taxonomer still correctly classifies 13.5% more reads compared to Kraken (59.5% vs. 46%) and also has a lower false positive rate (15.7% vs. 20.1%). Taxonomer also outperforms Kraken as regards taxon recovery rate (94.9% vs. 83%), and Taxonomer's false recovery rate is also lower (23.3% vs. We also examined the impact of read length and sequencing error rates upon classification accuracy. As would be expected, performance improved for all tools as a function of read lengths. We also found Taxonomer and Kraken to be more sensitive to sequencing errors than BLAST and the RDP Classifier. This is not surprising given their reliance upon exact k-mer matching. Nevertheless, these same analyses demonstrate that Taxonomer's nucleotide classification algorithm is tolerant to ~5% random error, with

Taxonomer achieving greater classification accuracies than Kraken. Figure 5.4d shows classification and recovery rates using Taxonomer's fungal database. As can be seen, the same general trends are seen in both Figure 5.4c and Figure 5.4d, demonstrating that Taxonomer's performance advantages are not restricted to bacterial classification.

Since quantifying microbial community composition is a frequent goal of metagenomics studies, we also compared Taxonomer's bacterial abundance estimates to those of the RDP Classifier using recently published 16S amplicon sequencing data and RNA-seq-based metagenomics (Figure 5.4e) (Lax et al., 2014; Subramanian et al., 2014). Taxonomer's abundance estimates are highly correlated with RDP's across taxonomic levels for all three datasets. Spearman Correlation coefficients (ρ) were 0.96 and 0.997 (order) and 0.858 and 0.826 (genus) for 16S amplicon data as well as 0.992 (order) and 0.955 (genus) for RNA-seq. However, Taxonomer's average analysis times were 260 to 440-fold faster (Figure 5.4e). Collectively, these benchmarks illustrate the important role of Taxonomer's classification databases and the power and speed of its classification algorithm.

Viral classification accuracy

Taxonomer uses reads from the 'viral' and 'unknown' bins for detection of viral and phage sequences via its Protonomer module (Figure 5.1a). To test classification performance, we compared Protonomer to two rapid protein search tools, RAPSearch2 (employed by SURPI) and DIAMOND (an ultrafast, BLAST-like protein search tool), using RNA-seq data from respiratory samples of 24 children with documented viral infections (Figure 5.2) (Buchfink et al., 2015; Y. Zhao et al., 2012). Protonomer demonstrated the best overall performance, being more sensitive (median 94.6%) than DIAMOND (90.5%) and more specific (90.7%) than RAPSearch2 (88.0%). As expected, sensitivity for all tools correlated with pairwise identities of viral genome to reference sequences with DIAMOND being most vulnerable to novel sequence polymorphisms. Of note, DIAMOND does not support joint analysis of paired sequencing reads. In this comparison, we used results of the mate pair with the lowest E-value rather than reconciling results of read mates, which likely results in optimistic performance estimates for DIAMOND. Protonomer is also the fastest of the three tools in classifying 10^4 to 10^6 reads/sample (Protonomer: 14 seconds; DIAMOND: 37 seconds in default and 46 seconds in sensitive modes; RAPSearch2: 343 seconds in default and 169 seconds in rapid modes).

We also used Taxonomer to analyze published RNA-seq data from three patients in whom viral pathogens of great public health significance were detected. These included a serum sample from a patient with hemorrhagic fever caused by a novel rhabdovirus (Bas Congo Virus, Figure 5.2d), a throat swab from a patient with avian influenza (H7N9 subtype, Figure 5.2e), and a plasma sample from a patient with Ebola virus (Figure 5.2f). Taxonomer detected the relevant viruses (or close relatives after removal of target sequences from the reference database) in all three cases, thus demonstrating the utility of Taxonomer for rapid virus detection and discovery in public health emergencies. Given its web-based deployment, this means that analysis results can be quickly shared and reviewed by experts, even across great geographic distances.

Human mRNA transcript profiling

Taxonomer also provides means for host response profiling, which is of growing interest for infectious diseases testing as well as quality control for cell lines and tissues where microbial contaminants may confound transcript expression profiles and lead to unsafe biologicals (Hudson et al., 2014; Mariotti et al., 2012). Taxonomer is the only ultrafast metagenomics tool with this capability. Taxonomer's default databases also include ERCC control sequences, allowing users to normalize transcript counts. We compared Taxonomer's expression profiles to those of standard transcript expression profiling tools (Sailfish, Cufflinks) (Patro et al., 2014; Trapnell et al., 2010). Taxonomer's quantification of synthetic reads and a commercially available RNA standard is accurate over a broad range of transcript abundance. Indeed, accuracy was intermediate between Sailfish's and Cufflink's (Figure 5.3A), demonstrating that Taxonomer provides state-of-the-art means for measuring transcript abundance.

To demonstrate utility of Taxonomer's capacity for simultaneous pathogen detection and transcript expression profiling, we analyzed RNA-seq data from respiratory samples of patients with influenza A virus infection (*n*=4) with varying abundance of host versus microbial RNA

(Figure 5.3b) and compared mRNA expression profiles to those of asymptomatic controls (*n*=40) (Anders & Huber, 2010; Jain et al., 2015). Influenza A virus could be detected in all samples by Taxonomer (see example in Figure 5.3c). Expression profiles for 17 host genes were significantly higher in influenza-positive patients (Figure 5.3d, examples in Figure 5.3f) and their expression profiles clearly differentiated cases from controls (Figure 5.3e). Gene ontology assignments for the top 50 genes demonstrated their involvement in recognition of pathogen-associated molecular patterns and antiviral host response (Figure 5.3g, Figure 5.3h). Most but not all of these genes are known to be differentially regulated in response to influenza virus or other viral infections *in vitro* or in peripheral blood of patients (Goujon et al., 2013; Haller, Staeheli, Schwemmle, & Kochs, 2015; X. Hu et al., 2013a; Zaas et al., 2013, 2009). Together, these results demonstrate the accuracy and power for discovery and a potential future diagnostic application of Taxonomer's combined pathogen detection and host response profiling.

Application of Taxonomer for microbial detection in a variety of real-world scenarios

In Figure 5.5, we show that Taxonomer can be used to detect previously unrecognized infectious diseases, to identify microbial contamination of stem cell cultures, and that it generates highly similar results with data from three commonly used next-generation sequencing platforms. We analyzed RNA-seq data from plasma of patients in whom Ebola virus disease was suspected but who had tested negative for Ebola virus (Gire et al., 2014). As was reported, Taxonomer detected HIV, Lassa virus, Enterovirus (typed by Taxonomer as Coxsackievirus), and GB virus C (data not shown). However, Taxonomer also detected previously unrecognized bacterial infections (*Chlamydophila psittaci*, *Elizabethkingia meningoseptica*), which may have caused the patients' symptoms (Figure 5.5a). *C. psittaci* is the agent of psittacosis, an uncommon zoonotic infection acquired from birds, that generally causes fever, headache, cough, and may also present with diarrhea. *E. meningoseptica* is a ubiquitous gram-negative bacterium that characteristically causes meningitis or sepsis in newborns but also immunocompromized adults. Given a high level of suspicion (as in an ongoing outbreak), these infections may have triggered testing for Ebola virus.

Taxonomer is not restricted to short reads, allowing reanalysis of the resulting contigs for greater classification sensitivity. Figure 5.5b shows Taxonomer results of 2,325 contigs generated from ‘viral’ and ‘unknown’ RNA-seq reads from a respiratory sample of a child with pneumonia (run time 6 seconds) (Jain et al., 2015). Four contigs were identified as unclassified members of the family Anelloviridae with 44%-60% predicted protein sequence identity to the most similar anellovirus. We also reanalyzed these data using Afterburner in combination with Protonomer, keeping track of resulting taxon assignments of each of the 239 reads in the anellovirus Trinity *de novo* assembly. Protonomer classified 19/239 of reads as anellovirus; Protonomer+Afterburner identified 89/239 reads as anellovirus. Protonomer did not misclassify any anellovirus reads, whereas Afterburner misclassified 110 of the anellovirus to other viral taxa. While probably not pathogenic, detection of this divergent Anellovirus demonstrates the power of Taxonomer for virus discovery.

Figure 5.5c shows RNA-seq data from induced pluripotent stem cell cultures with and without *Mycoplasma* contamination. Quality control of the RNA-seq results with Taxonomer immediately highlighted bacterial contamination (pie chart) and identified the organism as *M. yeatsii*.

Lastly, Taxonomer detected highly similar proportions of viral (influenza A, NP swab) and bacterial (*Mycoplasma pneumonia*, bronchoalveolar lavage*)* pathogens in respiratory tract samples subjected to 2 different library preparation methods and 3 different next-generation sequencing platforms (methods, Figure 5.5d). With each of the three platforms, >99% viral reads identified by Taxonomer were classified as influenza A virus. Proportion of bacterial 16S reads identified as *Mycoplasma pneumoniae* varied more (MiSeq 69.3%, HiSeq 65.9%, Ion Proton 30.5%). These results demonstrate the versatility of Taxonomer and how it can be used with a variety of sequencing instruments to detect previously missed pathogens and for quality control of expression profiling studies.

Discussion

In Taxonomer, we have created a tool that is fast, accurate, and capable of the gamut of analyses required to take full advantage of large and complex DNA/RNA-seq datasets for metagenomics. Taxonomer provides fast and effective means for read and contig classification, is substantially more accurate than the fastest available tools (Kraken or SURPI), and achieves accuracies on 16S amplicon data that closely approach the current standard, RDP. This is made possible by Taxonomer's comprehensive databases, its novel k-mer weighting approach, and its ability to carry out nucleotide and protein-based searches and classification within a single integrated algorithmic and visualization framework. Moreover Taxonomer is very fast, requiring only a few minutes to carry out its broad array of analyses. On the same typical HiSeq 2500 datasets, Taxonomer is hours faster than SURPI, days faster than RDP, and within minutes of the fastest published tool, Kraken, which only provides nucleotide classification.

We have produced a tool that is equally applicable to DNA and RNA-seq data, providing maximum flexibility for detection of known and unknown bacteria, fungi, as well as RNA and DNA viruses. Current estimates predict that the vast majority of bacteria, fungi, and viruses remain unknown and are thus not represented in reference sequence databases (Anthony et al., 2013; Koljalg et al., 2013; Rinke et al., 2013; Yarza et al., 2014). We have shown that 16S sequences (but not synthetic reads derived from other genomic targets) from the same unrepresented bacteria are almost always correctly binned by Taxonomer (but not erroneously classified), highlighting the advantages of Taxonomer's marker gene-based approach both for discovery of novel organisms and for avoiding misclassifications pitfalls (Afshinnekoo et al., 2015). Integrated means to search and classify in nucleotide and protein space improves sensitivity, especially for detection of viruses. This is due to high mutation rates and high sequence diversity in many viral phyla, rendering sequence homologies more readily detectable at the protein level rather than at the nucleotide level.

Taxonomer's integrated framework means that microorganisms can be classified in nucleotide or protein space using the same k-mer weighting-based approach and classification algorithm. The result is greater tolerance for sequencing errors, better sensitivity, more accurate abundance estimates, and execution times that exceed even those of the fastest published protein search tools. This speed and breath of functionality is crucial, as many clinical samples contain complex mixtures of bacterial, fungal, and viral taxa. We have successfully demonstrated the use of Taxonomer in real-world scenarios to identify a diverse set of known viruses (respiratory viruses, HIV, Lassa virus, Coxsackievirus, GB virus C), unexpected viruses (Bas Congo Virus, avian influenza A virus H7N9), and unrecognized bacteria and viruses in previously test-negative patients (Anellovirus, *Chlamydophila psittaci*, *Elizabethkingia meningoseptica*).

Taxonomer also provides automatic means to classify host gene expression using the same integrated methodology, a functionality that enables new analysis modalities for ultrafast metagenomics. For example, the simultaneous identification of viral pathogens and characterization of host transcriptional responses provides information that can be leveraged for greater diagnostic power and precision. Similar results have been obtained using blood, but our demonstration of Taxonomer's ability to rapidly identify children with influenza virus infection directly from upper respiratory tract specimens using only their (own) mucosal gene expression profiles has important implications for diagnosis and discovery (X. Hu et al., 2013a; Zaas et al., 2009, 2013). Other, equally novel applications are also possible. Examples include differentiating true infections from asymptomatic carriage based on the host response, characterizing chronic infections in immunocompromised patients, and real-time monitoring of the impacts of antimicrobial treatment in conjunction with host-transcriptional responses, all of which hold much promise for improved patient care, antimicrobial stewardship, and epidemiological investigations.

We further demonstrate how Taxonomer is used to address a crucial, widespread unrecognized microbial contamination or infection issue in RNA-seq studies, which can heavily confound transcriptional responses of cells in culture or from biopsy (Olarerin-George & Hogenesch, 2015). In addition, sample contamination by exogenous sequences directly or through their presence in commonly used laboratory reagents and kits can lead to erroneous genome assemblies and disease associations, further highlighting the need for thorough quality control of sequencing reads (Cantalupo, Katz, & Pipas, 2015; Merchant, Wood, & Salzberg, 2014; Naccache et al., 2013; Rosseel, Pardon, De Clercq, Ozhelvaci, & Van Borm, 2014; Smuts, Kew, Khan, & Korsman, 2014; Strong et al., 2014). This is of particular concern when source DNA or RNA is of low concentration, such as is the case with single-cell sequencing studies (Lusk, 2014). Clearly, Taxonomer's ability to simultaneously quantify transcriptional responses and to monitor DNA and RNA-seq datasets for signs of infection and contamination will benefit scientific and diagnostic applications alike. Lastly, metagenomic sequencing data are usually purged of host sequences prior to deposition in public sequence databases to guarantee anonymity of patients (Rotmistrovsky & Agarwala, 2011; Sherry, 2011). During analysis of some such sequences with Taxonomer, varying numbers of human sequences were detected, suggesting that the Binner module is more effective at detecting (and removing) host-derived sequences than currently used tools (Gire et al., 2014). Therefore, screening of metagenomics datasets with Taxonomer prior to their submission could improve protection of study subjects' privacy.

Finally, with Taxonomer, we have sought to democratize these analyses by providing a fast interactive web service based upon the iobio visualization toolkit (Miller et al., 2014). As our analyses of RNA-seq data from patients harboring viral pathogens of great public health significance demonstrate, Taxonomer provides effective means for rapid virus detection for patient care and discovery in public health emergencies. The ability to conveniently upload and rapidly analyze samples from personal computers and mobile devices via the Taxonomer web-portal means that analysis results can be quickly shared and reviewed by experts, even across great geographic distances enhancing collaborations and facilitating public health responses. As costs and turn-around times for high-throughput sequencing continue to fall, Taxonomer will enable a rapidly growing number of diagnostic laboratories with access to sequencing instruments to analyze data in a meaningful timeframe without having to invest in computational infrastructure or bioinformatics expertise.

References

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., … Mason, C. E. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Systems*, *1*(1), 1–15. doi:http://dx.doi.org/10.1016/j.cels.2015.01.001

Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica*, *96*(5), 644–647. doi:10.1111/j.1651-2227.2006.00178.x

Altschul, S., Gish, W., & Miller, W. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology,* 215(3), 403-410. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022283605803602

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrelio, C. M., Solovyov, A., … Lipkin, W. I. (2013). A strategy to estimate unknown viral diversity in mammals. *mBio*, *4*(5), e00598–13. doi:10.1128/mBio.00598-13

Audano, P., & Vannberg, F. (2014). KAnalyze: A fast versatile pipelined K-mer toolkit. *Bioinformatics*, *30*(14), 2070–2072. doi:10.1093/bioinformatics/btu152

Borozan, I., Watt, S. N., & Ferretti, V. (2013). Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. *PLoS One*, *8*(10), e76935. doi:10.1371/journal.pone.0076935

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. doi:10.1038/nmeth.3176

Caliendo, A. M., Gilbert, D. N., Ginocchio, C. C., Hanson, K. E., May, L., Quinn, T. C., … Infectious Diseases Society of, A. (2013). Better tests, better care: Improved diagnostics for infectious diseases. *Clinical Infectious Diseases*, *57 Suppl 3*, S139–70. doi:10.1093/cid/cit578

Cantalupo, P. G., Katz, J. P., & Pipas, J. M. (2015). HeLa nucleic acid contamination in The Cancer Genome Atlas leads to the misidentification of HPV18. *Journal of Virology*, *89*(8), 4051–4057. doi:10.1128/JVI.03365-14

Chiu, C. Y. (2013). Viral pathogen discovery. *Current Opinion in Microbiology*, *16*(4), 468–478. doi:10.1016/j.mib.2013.05.001

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633–42. doi:10.1093/nar/gkt1244

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. doi:10.1128/AEM.03006-05

Dickson, R. P., Martinez, F. J., & Huffnagle, G. B. (2014). The role of the microbiome in exacerbations of chronic lung diseases. *Lancet*, *384*(9944), 691–702. doi:10.1016/S0140-6736(14)61136-3

Firth, C., Bhat, M., Firth, M. A., Williams, S. H., Frye, M. J., Simmonds, P., … Lipkin, W. I. (2014). Detection of zoonotic pathogens and characterization of novel viruses carried by commensal Rattus norvegicus in New York City. *mBio*, *5*(5), e01933–14. doi:10.1128/mBio.01933-14

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology*, *12*, 69. doi:doi:10.1186/s12915-014-0069-1

Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., … Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, *345*(6202), 1369–1372. doi:10.1126/science.1259657

Goujon, C., Moncorge, O., Bauby, H., Doyle, T., Ward, C. C., Schaller, T., … Malim, M. H. (2013). Human MX2 is an interferon-induced post-entry inhibitor of HIV-1 infection. *Nature*, *502*(7472), 559–562. doi:10.1038/nature12542

Graf, E. H. (2015). Evaluation of metagenomics for the detection of respiratory viruses directly from clinical samples. *Under Review*.

Grard, G., Fair, J. N., Lee, D., Slikas, E., Steffen, I., Muyembe, J. J., … Leroy, E. M. (2012). A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathogens*, *8*(9), e1002924. doi:10.1371/journal.ppat.1002924

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R., & Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, *40*(20), 10073–10083. doi:10.1093/nar/gks666

Haller, O., Staeheli, P., Schwemmle, M., & Kochs, G. (2015). Mx GTPases: Dynamin-like antiviral machines of innate immunity. *Trends in Microbiology*, *23*(4), 154–163. doi:10.1016/j.tim.2014.12.003

Hu, X., Yu, J., Crosby, S. D., & Storch, G. A. (2013a). Gene expression profiles in febrile children with defined viral and bacterial infection. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(31), 12792–12797. doi:10.1073/pnas.1302968110

Hu, Y., Lu, S., Song, Z., Wang, W., Hao, P., Li, J., … Yuan, Z. (2013). Association between adverse clinical outcome in human disease caused by novel influenza A H7N9 virus and sustained viral shedding and emergence of antiviral resistance. *Lancet*, *381*(9885), 2273–2279. doi:10.1016/S0140-6736(13)61125-3

Hudson, L. L., Woods, C. W., & Ginsburg, G. S. (2014). A novel diagnostic approach may reduce inappropriate antibiotic use for acute respiratory infections. *Expert Review of Anti-Infective Therapy*, *12*(3), 279–282. doi:10.1586/14787210.2014.881717

Human Microbiome Project Consortium. (2012). A framework for human microbiome research. *Nature*, *486*(7402), 215–221.

Jain, S., Williams, D. J., Arnold, S. R., Ampofo, K., Bramley, A. M., Reed, C., … Finelli, L. (2015). Community-Acquired Pneumonia Requiring Hospitalization among U.S. Children. *New England Journal of Medicine*, *372*(9), 835–845. doi:10.1056/NEJMoa1405870

Koljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., … Larsson, K. H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277. doi:10.1111/mec.12481

Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., … Gilbert, J. A. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, *345*(6200), 1048–1052. doi:10.1126/science.1254529

Lipkin, W. I. (2013). The changing face of pathogen discovery and surveillance. *Nature Reviews. Microbiology*, *11*(2), 133–41. doi:10.1038/nrmicro2949

Louis, P., Hold, G. L., & Flint, H. J. (2014). The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, *12*(10), 661–672. doi:10.1038/nrmicro3344

Lusk, R. W. (2014). Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. *PLoS One*, *9*(10), e110808. doi:10.1371/journal.pone.0110808

Mariotti, E., D'Alessio, F., Mirabelli, P., Di Noto, R., Fortunato, G., & Del Vecchio, L. (2012). Mollicutes contamination: A new strategy for an effective rescue of cancer cell lines. *Biologicals*, *40*(1), 88–91. doi:10.1016/j.biologicals.2011.10.006

Mayer, E. A., Tillisch, K., & Gupta, A. (2015). Gut/brain axis and the microbiota. *The Journal of Clinical Investigation*, *125*(3), 926–938. doi:10.1172/JCI76304

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., … Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610–618. doi:10.1038/ismej.2011.139

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, *2*, e675. doi:10.7717/peerj.675

Miller, C. A., Qiao, Y., DiSera, T., D'Astous, B., & Marth, G. T. (2014). bam.iobio: a web-based, real-time, sequence alignment file inspector. *Nature Methods*, *11*(12), 1189. doi:10.1038/nmeth.3174

Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., … Chiu, C. Y. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research*, *24*(7), 1180–1192. doi:10.1101/gr.171934.113

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., … Chiu, C. Y. (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *Journal of Virology*, *87*(22), 11966–11977. doi:10.1128/JVI.02323-13

Olarerin-George, A. O., & Hogenesch, J. B. (2015). Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Research*, *43*(5), 2535–2542. doi:10.1093/nar/gkv136

Patro, R., Mount, S. M., & Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature Biotechnology*, *32*(5), 462–4. doi:10.1038/nbt.2862

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. doi:10.1038/nature12352

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., & Van Borm, S. (2014). False-positive results in metagenomic virus discovery: A strong case for follow-up diagnosis. *Transboundary and Emerging Diseases*, *61*(4), 293–299. doi:10.1111/tbed.12251

Rotmistrovsky, K., & Agarwala, R. (2011). BMTagger: Best match tagger for removing human reads from metagenomics datasets. *Bioinformatics*, *Unpublished*.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., … Ye, J. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *38*(Database issue), D5–16. doi:10.1093/nar/gkp967

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., & Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, *15*(6), 1882–1899. doi:10.1111/1462-2920.12086

Sherrard, L. J., Tunney, M. M., & Elborn, J. S. (2014). Antimicrobial resistance in the respiratory microbiota of people with cystic fibrosis. *Lancet*, *384*(9944), 703–713. doi:10.1016/S0140-6736(14)61137-5

Sherry, S. (2011). Human Sequence Removal, National Center or Biotechnology Information. http://www.hmpdacc.org/doc/HumanSequenceRemoval_SOP.pdf. doi:http://www.hmpdacc.org/doc/HumanSequenceRemoval_SOP.pdf

Smuts, H., Kew, M., Khan, A., & Korsman, S. (2014). Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *Journal of Virology*, *88*(2), 1398. doi:10.1128/JVI.03206-13

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., … Flemington, E. K. (2014). Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. *PLoS Pathogens*, *10*(11), e1004437. doi:10.1371/journal.ppat.1004437

Subramanian, S., Huq, S., Yatsunenko, T., Haque, R., Mahfuz, M., Alam, M. A., … Gordon, J. I. (2014). Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature*, *510*(7505), 417–421. doi:10.1038/nature13421

The Greengenes Database. http://greengenes.secondgenome.com. (n.d.). Retrieved from http://greengenes.secondgenome.com

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*(9), 1105–1111. doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. doi:10.1038/nbt.1621

“UNITE.” (2014). UNITE. http://unite.ut.ee. Retrieved from http://unite.ut.ee

Wilson, M. R., Naccache, S. N., Samayoa, E., Biagtan, M., Bashir, H., Yu, G., … Chiu, C. Y. (2014). Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *New England Journal of Medicine*, *370*(25), 2408–2417. doi:10.1056/NEJMoa1401268

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. doi:10.1186/gb-2014-15-3-r46

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F. O., Ludwig, W., Schleifer, K. H., … Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645. doi:10.1038/nrmicro3330

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., … Glockner, F. O. (2014). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, *42*(Database issue), D643–8. doi:10.1093/nar/gkt1209

Zaas, A. K., Burke, T., Chen, M., McClain, M., Nicholson, B., Veldman, T., … Ginsburg, G. S. (2013). A host-based RT-PCR gene expression signature to identify acute respiratory viral infection. *Science Translational Medicine*, *5*(203), 203ra126. doi:10.1126/scitranslmed.3006280

Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero 3rd, A. O., Lucas, J., … Ginsburg, G. S. (2009). Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host & Microbe*, *6*(3), 207–217. doi:10.1016/j.chom.2009.07.006

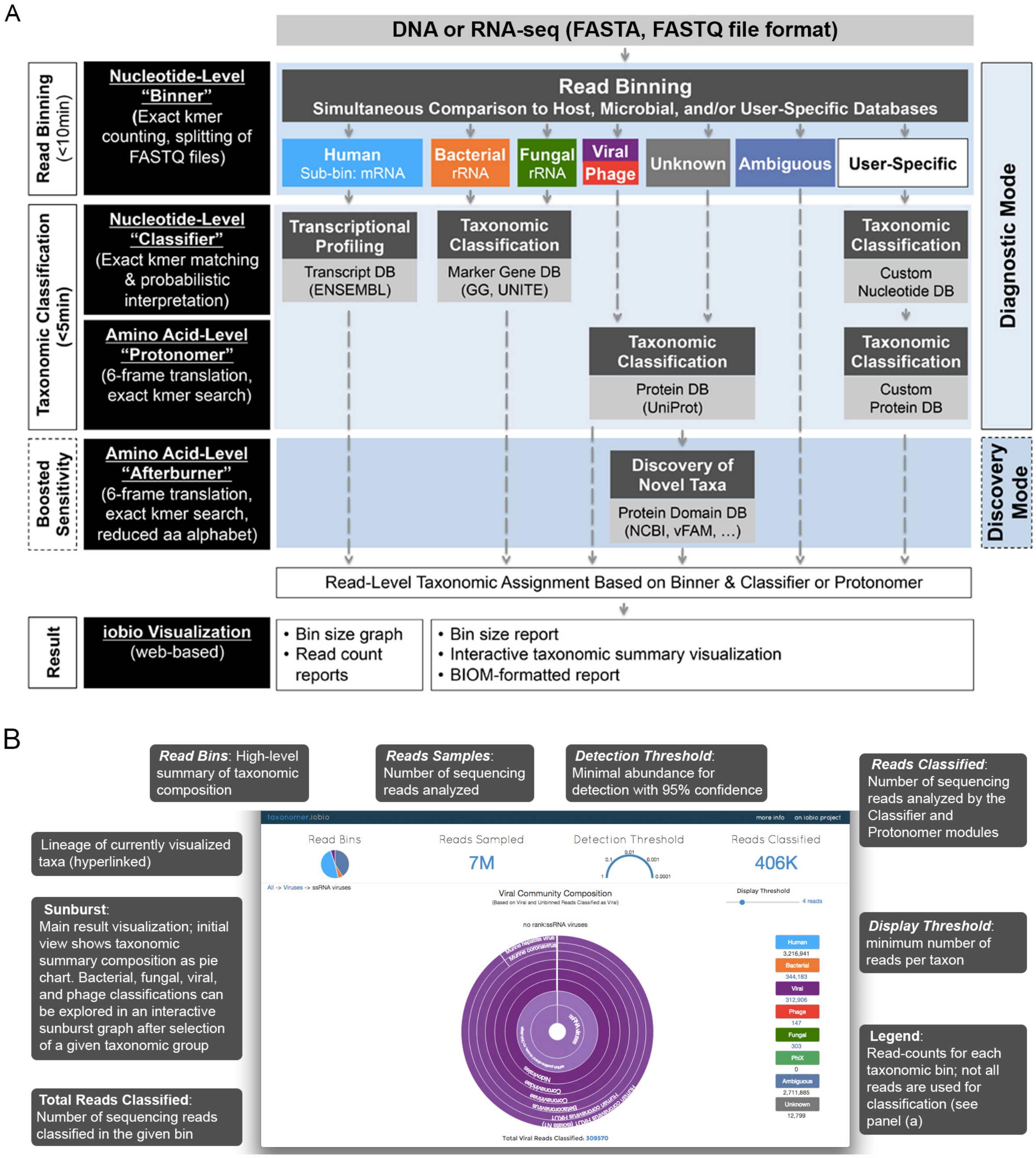
Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., … Sittler, T. (2011). Faster and more accurate sequence alignment with SNAP. *arXiv.org*, arXiv:1111.5572.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, *7*(1-2), 203–214. doi:10.1089/10665270050081478

Zhao, L. (2013). The gut microbiota and obesity: From correlation to causality. *Nature Reviews Microbiology*, *11*(9), 639–647. doi:10.1038/nrmicro3089

Zhao, Y., Tang, H., & Ye, Y. (2012). RAPSearch2: A fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, *28*(1), 125–126. doi:10.1093/bioinformatics/btr595

Figure 5.1. Overview of Taxonomer architecture and user interface. (a) Taxonomer's architecture. Raw FASTA, FASTQ, or SRA files (with or without gzip compression) are the input for Taxonomer. For paired-end data, mate pairs are analyzed jointly. Taxonomer consists of four main modules. The 'Binner' module categorizes ('bins') reads into broad taxonomic groups (host and microbial) followed by comprehensive microbial and host gene expression profiling at the nucleotide ('Classifier' module) or amino acid-level ('Protonomer' and 'Afterburner' modules). Normalized host gene expression (gene-level read counts) and microbial profiles. Read subsets can be downloaded for custom downstream analyses **(b) Taxonomer web-service.** To further remove barriers for academic and clinical adoption of metagenomics, we developed a web interface for Taxonomer that allows users to stream sequencing read files (stored locally or http accessibly) to the analysis server and interactively visualize results in real-time. Main features are described in grey boxes. Taxonomic classification of bacteria, fungi, and viruses is visualized as a sunburst graph (center), in which the size of a given slice represents the relative abundance at the read level. Taxonomic ranks are shown hierarchically with the highest rank in the center of the graph. Sequences that cannot be classified to the species level, either because they are shared between taxa or represent novel microorganisms, are collapsed to the lowest common ancestor and shown as part of slices that terminate at higher taxonomic ranks like genus or family.


A
DNA or RNA-seq (FASTA, FASTQ file format)
Read Binning (<10min)
Nucleotide-Level "Binner" (Exact kmer counting, splitting of FASTQ files)
Read Binning
Simultaneous Comparison to Host, Microbial, and/or User-Specific Databases
Human Sub-bin: mRNA
Bacterial rRNA
Fungal rRNA
Viral Phage
Unknown
Ambiguous
User-Specific
Diagnostic Mode
Taxonomic Classification (<5min)
Nucleotide-Level "Classifier" (Exact kmer matching & probabilistic interpretation)
Amino Acid-Level "Protonomer" (6-frame translation, exact kmer search)
Transcriptional Profiling
Transcript DB (ENSEMBL)
Taxonomic Classification
Marker Gene DB (GG, UNITE)
Taxonomic Classification
Custom Nucleotide DB
Taxonomic Classification
Protein DB (UniProt)
Taxonomic Classification
Custom Protein DB
Boosted Sensitivity
Amino Acid-Level "Afterburner" (6-frame translation, exact kmer search, reduced aa alphabet)
Discovery of Novel Taxa
Protein Domain DB (NCBI, vFAM, …)
Discovery Mode
Read-Level Taxonomic Assignment Based on Binner & Classifier or Protonomer
Result
iobio Visualization (web-based)
• Bin size graph
• Read count reports
• Bin size report
• Interactive taxonomic summary visualization
• BIOM-formatted report
B
taxonomer.iobio
more info
an iobio project
Read Bins
Reads Sampled
7M
Detection Threshold
0.01
0.1
0.001
1
0.0001
Reads Classified
406K
All -> Viruses -> ssRNA viruses
Viral Community Composition
(Based on Viral and Unbinned Reads Classified as Viral)
Display Threshold
4 reads
no rank:ssRNA viruses
Murine hepatitis virus
Murine coronavirus
ssRNA viruses
ssRNA positive-strand viruses, no DNA stage
Nidovirales
Coronaviridae
Coronavirinae
Betacoronavirus
Human coronavirus HKU1
Human coronavirus HKU1 (isolate N1)
Human
3,216,941
Bacterial
344,183
Viral
312,906
Phage
147
Fungal
303
PhiX
0
Ambiguous
2,711,885
Unknown
12,799
Total Viral Reads Classified: 309570

Read Bins: High-level summary of taxonomic composition

Reads Samples: Number of sequencing reads analyzed

Detection Threshold: Minimal abundance for detection with 95% confidence

Reads Classified: Number of sequencing reads analyzed by the Classifier and Protonomer modules

Lineage of currently visualized taxa (hyperlinked)

Sunburst: Main result visualization; initial view shows taxonomic summary composition as pie chart. Bacterial, fungal, viral, and phage classifications can be explored in an interactive sunburst graph after selection of a given taxonomic group

Total Reads Classified: Number of sequencing reads classified in the given bin

Display Threshold: minimum number of reads per taxon

Legend: Read-counts for each taxonomic bin; not all reads are used for classification (see panel (a)

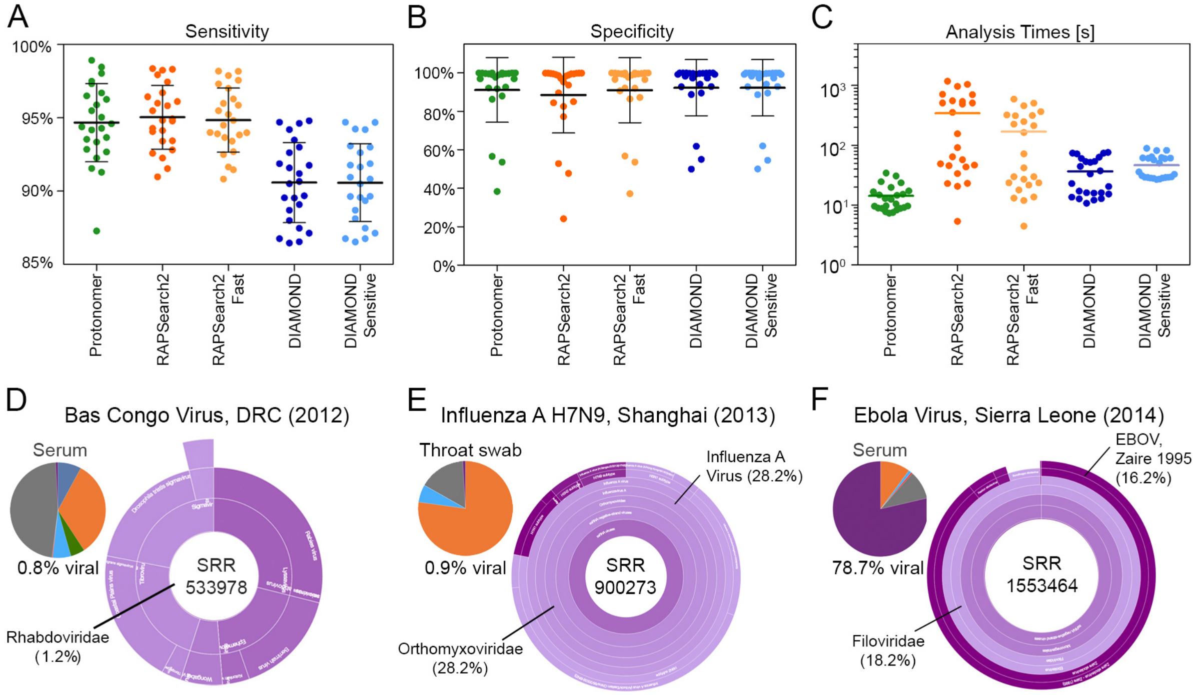
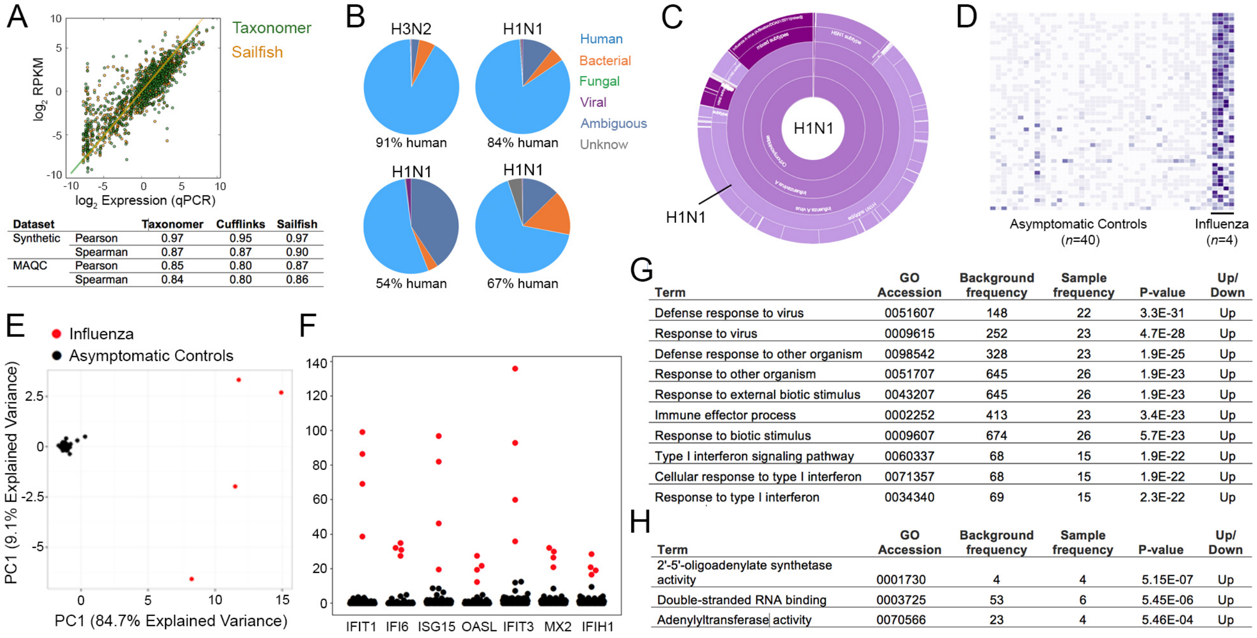


Figure 5.2. **Performance of the 'Classifier' module for bacterial and fungal classification, and bacterial community profiling. (a)** Taxonomer provides superior sensitivity and specificity for read-level bacterial classification compared to two other rapid classification tools SURPI and Kraken when using each tool's default settings and databases: nt (SURPI), RefSeq (Kraken), and Greengenes 99% OTU (Taxonomer). Results for SURPI are based on correct identification by either (dark bar) or both (light bar) read mates. **(b)** Of the three commonly-used reference databases RefSeq (*n*=210,627; 5,242 bacterial genomes), Greengenes 99% OTU (*n*=203,452), and RDP (*n*= 2,929,433), Taxonomer provides greatest read-level (top) and taxon-level (bottom, that gives the percentage of bacterial species identified) sensitivity for bacterial classification at only a moderate decrease in specificity when using the Greengenes database compared to the RDP and RefSeq databases (simulated 16S rDNA as in panel a). Because of its large size and greater completeness, the RDP database provides the greatest species-level specificity at the tradeoff of sensitivity. For ease of reference, the top right-most column is repeated from panel a. **(c)** Bacterial classification accuracy of Taxonomer is similar to the RDP Classifier and superior to Kraken at the read-level (top) and taxon-level (bottom, all using the Greengenes database). Given the applied criteria, BLAST is less sensitive but more specific. **(d)** Taxonomer also performs similar to the RDP Classifier and better than Kraken for classification of synthetic fungal internal transcribed spacer (ITS) sequences at the read-level (top) and taxon-level (bottom). **(e)** Taxonomer classifies bacterial 16S rRNA reads at >200-fold increased speed compared to the RDP Classifier (times for 1 CPU, multithreading not available for RDP Classifier) while providing highly comparable bacterial community profiles when using 16S rRNA gene amplicon sequencing and shotgun metagenomics. Spearman correlation coefficients (ρ) of abundance estimates are shown for Taxonomer and the RDP Classifier at the order and genus-levels using the Greengenes 99% OTU reference database.



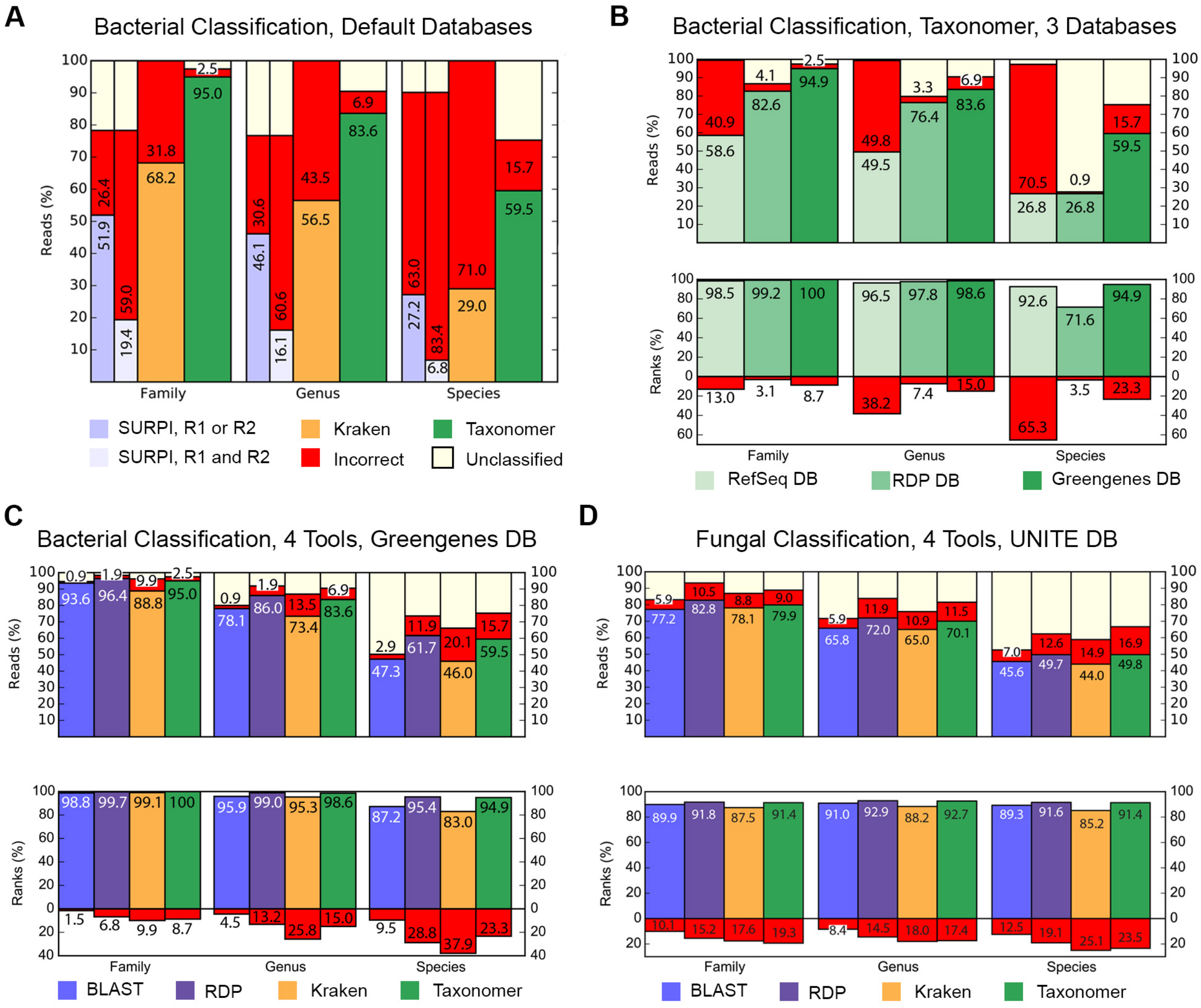
Dataset		Taxonomer	Cufflinks	Sailfish
Synthetic	Pearson	0.97	0.95	0.97
	Spearman	0.87	0.87	0.90
MAQC	Pearson	0.85	0.80	0.87
	Spearman	0.84	0.80	0.86

Term	GO Accession	Background frequency	Sample frequency	P-value	Up/ Down
Defense response to virus	0051607	148	22	3.3E-31	Up
Response to virus	0009615	252	23	4.7E-28	Up
Defense response to other organism	0098542	328	23	1.9E-25	Up
Response to other organism	0051707	645	26	1.9E-23	Up
Response to external biotic stimulus	0043207	645	26	1.9E-23	Up
Immune effector process	0002252	413	23	3.4E-23	Up
Response to biotic stimulus	0009607	674	26	5.7E-23	Up
Type I interferon signaling pathway	0060337	68	15	1.9E-22	Up
Cellular response to type I interferon	0071357	68	15	1.9E-22	Up
Response to type I interferon	0034340	69	15	2.3E-22	Up

Term	GO Accession	Background frequency	Sample frequency	P-value	Up/ Down
2'-5'-oligoadenylate synthetase activity	0001730	4	4	5.15E-07	Up
Double-stranded RNA binding	0003725	53	6	5.45E-06	Up
Adenylyltransferase activity	0070566	23	4	5.46E-04	Up

Figure 5.3. Performance characteristics of the 'Classifier' module for host transcript expression profiling. (a) Published RNA-seq data from a commercially available RNA standard (MAQC) were analyzed by Taxonomer, Sailfish, and Cufflinks and estimated transcript expression was compared to data obtained by quantitative PCR (qPCR). Gene-level Pearson and Spearman correlation coefficients for RNA-seq versus qPCR were 0.85 and 0.84 for Taxonomer, 0.87 and 0.86 for Sailfish, and 0.80 and 0.80 for Cufflinks, respectively. (**b**) Application of Taxonomer to metagenomic RNA-seq data from routine respiratory samples from patients with influenza infection (*n*=4). **(c)** Panel C shows classification of viral sequencing reads by Protonomer and typing of this strain as influenza A(H1N1)pdm09 (top right sample from panel A). **(d)** Differential gene-level mRNA expression profiles from 4 patients with influenza A virus compared to asymptomatic controls (*n*=40; top 50 differentially expressed genes are shown). Expression profiles for 17 genes were significantly higher in influenza-positive patients. **(e)** Expression profiles for the 17 most differentially expressed genes differentiate cases from controls (principal component analysis, PC1 and PC2 explaining 93.8% of the total variance). **(f)** Normalized expression levels for individual patients of seven of the top 17 genes. Gene ontology assignments for enrichment of biological processes **(g)** and molecular functions **(h)** are shown.

Figure 5.4. **Performance of the 'Classifier' module for bacterial and fungal classification, and bacterial community profiling. (a)** Taxonomer provides superior sensitivity and specificity for read-level bacterial classification compared to two other rapid classification tools SURPI[23] and Kraken when using each tool's default settings and databases: nt (SURPI), RefSeq (Kraken), and Greengenes 99% OTU (Taxonomer). Results for SURPI are based on correct identification by either (dark bar) or both (light bar) read mates. **(b)** Of the three commonly-used reference databases RefSeq (*n*=210,627; 5,242 bacterial genomes), Greengenes 99% OTU (*n*=203,452), and RDP (*n*= 2,929,433), Taxonomer provides greatest read-level (top) and taxon-level (bottom, which is the percentage of bacterial species identified) sensitivity for bacterial classification at only a moderate decrease in specificity when using the Greengenes database compared to the RDP and RefSeq databases (simulated 16S rDNA as in panel a). Because of its large size and greater completeness, the RDP database provides the greatest species-level specificity at the tradeoff of sensitivity. For ease of reference, the top right-most column is repeated from panel a. **(c)** Bacterial classification accuracy of Taxonomer is similar to the RDP Classifier and superior to Kraken at the read-level (top) and taxon-level (bottom, all using the Greengenes database). Given the applied criteria, BLAST is less sensitive but more specific. **(d)** Taxonomer also performs similar to the RDP Classifier and better than Kraken for classification of synthetic fungal internal transcribed spacer (ITS) sequences at the read-level (top) and taxon-level (bottom). **(e)** Taxonomer classifies bacterial 16S rRNA reads at >200-fold increased speed compared to the RDP Classifier (times for 1 CPU, multithreading not available for RDP Classifier) while providing highly comparable bacterial community profiles when using 16S rRNA gene amplicon sequencing and shotgun metagenomics. Spearman correlation coefficients (ρ) of abundance estimates are shown for Taxonomer and the RDP Classifier at the order and genus-levels using the Greengenes 99% OTU reference database.



E

Sample	Approach	Platform, Reads	Reads/Sample	Time/Sample [s] Taxonomer	RDP	Rank	ρ
Stool [12]	16S (V4) amplicon	HiSeq (1x150bp)	$1.6x10^4$	0.7 min	311 min	Order	0.960
						Genus	0.858
Environment, human, pets[13]	16S (V4) amplicon	MiSeq (2x150bp)	$7.3x10^4$	1.8 min	644 min	Order	0.997
						Genus	0.826
Respiratory	RNA-Seq	HiSeq (2x100 bp)	$1.6x10^6$	27.4 min	7,245 min	Order	0.992
						Genus	0.955

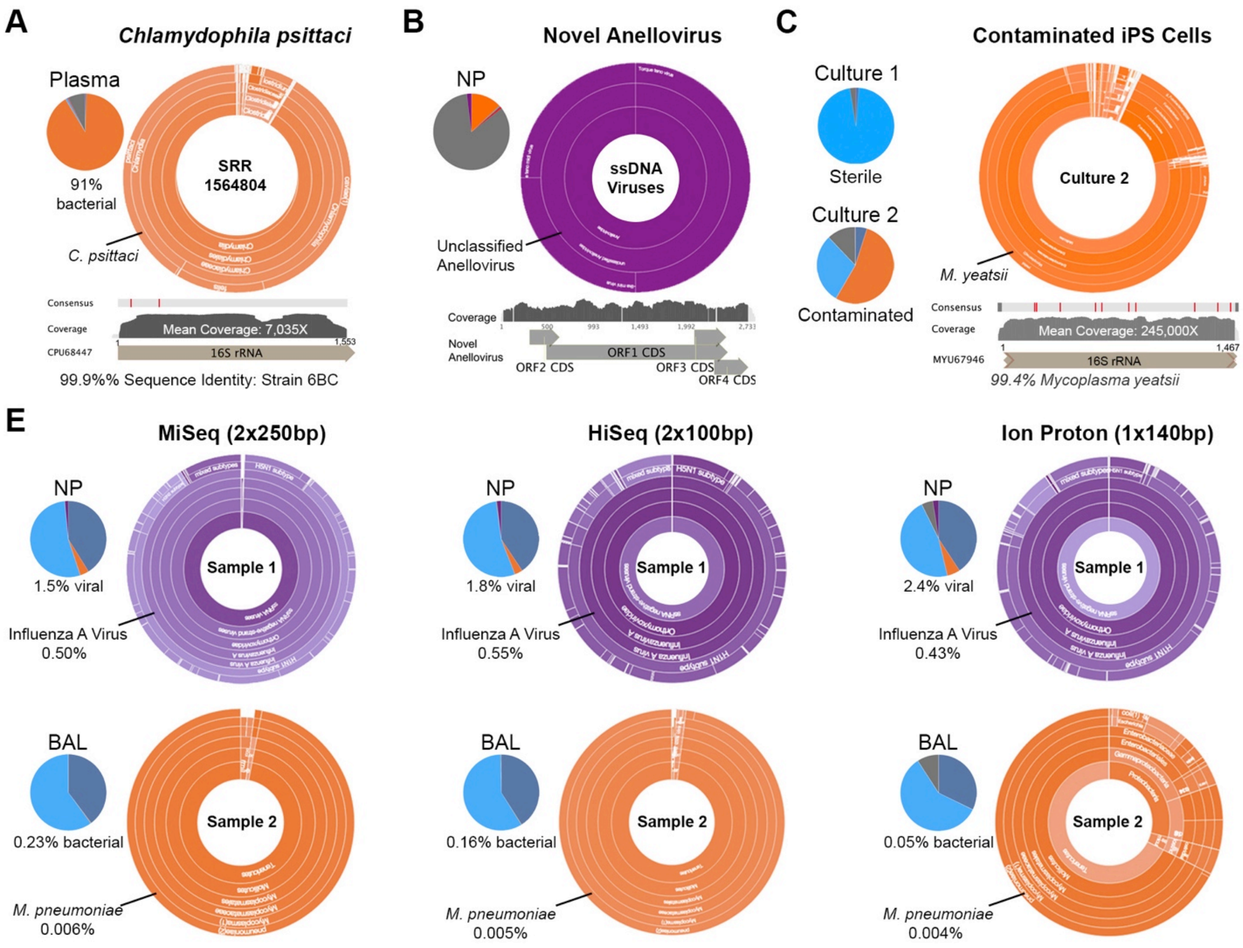
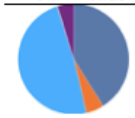
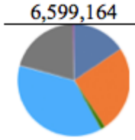
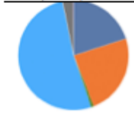


Figure 5.5. Sample applications of Taxonomer. (a) Taxonomer detected a previously unrecognized *Chlamydophila psittaci* infection (psittacosis), in plasma from a patient with suspected Ebola virus disease in Sierra Leone (SRR1564804)[32]. The 16S rRNA gene was covered a mean of 7,035-fold with the consensus 16S rRNA sequence from this isolate sharing 99.9% identity with the type strain (6BC, ATCC VR-125, CPU68447) enabling reliable identification[75]. Positions of 2 single nucleotide polymorphisms are highlighted in red. **(b)** Taxonomer detected a novel Anellovirus in a nasopharyngeal swab. Pie chart and sunburst show contig-level classification (*de novo* assembly with Trinity[36]). Mapping reads back to a manually-constructed viral consensus genome sequence showed x-fold coverage, 68.5% pairwise nucleotide-level identity and 44%-60% predicted protein identity with TTV-like mini virus isolate LIL-y1 **(**EF538880.1). **(c)** Identification of *Mycoplasma yeatsii* contamination in RNA-seq data from cultured iPS cell (right) compared to non-contaminated iPS cell culture (left) based on read binning (top). High expression of rRNA is demonstrated by 32% of RNA-Seq reads mapping to the *M. yeatsii* 16S rRNA gene (245,000X coverage, 99.4% sequence identity with type strain GIH (MYU67946). **(d)** Taxonomer is compatible with different sequencing protocols, recovering similar proportions of viral (influenza A, 0.43% to 0.55% of all reads) and bacterial (*Mycoplasma pneumoniae*, 16S rRNA sequences representing 0.004% to 0.006% of all reads) pathogen sequences when sequencing samples on 3 commonly-used sequencers with 2 different library preparation methods. Samples were known to be positive for influenza A(H1N1)pdm09 and *M. pneumoniae* based on diagnostic PCR test.

Table 5.1. Processing time of Taxonomer compared to rapid classification pipelines SURPI and Kraken. Five RNA-Seq samples generated from nasal specimens with varying degrees of taxonomic composition illustrate the effect on pipeline speeds. (Human-blue; Bacteria-orange; Fungal-green; Virus-red; other-yellow; unclassified-grey).

Sample Composition, Total Reads	Pathogen	Application	Subtraction	Binning	Classification	Protein Search	Total Time	% Reads Classified
 6,599,164	**HCoV**	Taxonomer	-	5m	22s	10s	5.5m	99%
		Kraken	-	-	1.5m	-	1.5m	99%
		SURPI	3.3m	-	74m	15m	92m	98%
 7,542,552	**Influenza A virus**	Taxonomer	-	8m	40s	30s	9.2m	77%
		Kraken	-	-	1.5m	-	1.5m	66%
		SURPI	9.8m	-	208m	18m	236m	78%
 6,252,311	**HMPV**	Taxonomer	-	5.2m	56s	10s	6.3m	97%
		Kraken	-	-	1.3m	-	1.3m	93%
		SURPI	56m	-	648m	24m	728m	95%

CHAPTER 6

CONCLUSIONS

Computational approaches to biological data

Experiments in the biological sciences increasingly are producing datasets large enough that manual analyses are impossible. This increase in data presents a lot of scientific opportunity as well as challenges computationally in the analysis. In my dissertation, I have presented effective computational solutions to analyze image data, prioritize human genetic variants, and to comprehensively analyze metagenomic data.

Image analysis

Modern microscopes can produce thousands on high quality images in a relatively short amount of time. Thus, automated image analysis has a large impact potential in many of the biological sciences. There are many excellent open source image analysis packages for the Python programming language that provide implementations of standard image analysis functions. Using Python and open source image analysis packages, I created an open source image analysis pipeline, ImagePlane, to process images of *S. mediterranea* (details of the pipeline are given in chapter 2) (Flygare, Campbell, Ross, Moore, & Yandell, 2013). Chapter 3 demonstrates the application of image analysis to analyze muscle fiber size with another open source image analysis pipeline I created, MuscleQNT, which is also written in Python. MuscleQNT includes functionality to analyze images of stained muscle cross sections, create histograms of muscle fiber sizes, and perform statistical tests to find biologically relevant differences between mutant and control animals. To my knowledge, when created, these image analysis pipelines provided unique analysis ability in their particular application domains. MuscleQNT has enabled analyses that have been published.

These pipelines demonstrate the power of combining existing image analysis and statistical libraries into tools that enable directed analyses that would otherwise be incomplete or impossible. I believe that scientists performing or directing the analysis of images need at least a basic understanding of core image analysis procedures like image thresholding, erosion and dilation methods, and feature size and location quantification. An understanding of these methods will enable an increased ability to craft and interpret the analyses specific to the data and experiment at hand. Chapter 3 is an excellent example of crafted image analysis together with statistical / graphical analysis for the specific experiment.

Human variant prioritization

As sequencing costs have dropped, the amount of human sequencing has skyrocketed, which has resulted in tens of millions of known variants in public databases (the NCBI's dbSNP database contains more than 100 million human variants). Given all this known variation, perhaps the most important question to be asked is how to rank variants according to their relative risk in human disease. Given any particular variant, how do we determine how likely it is to contribute to human disease? This is the task of variant prioritization. There have been many methods published as solutions to human variant prioritization; however, all of them suffer from significant limitations (Katsonis et al., 2014; Kircher et al., 2014). Perhaps the greatest limitation of the majority of these tools is they are not able to prioritize all variants – instead, they prioritize some small subset like variants that induce nonsynonymous changes. To my knowledge, CADD and VVP are the only tools that can prioritize nearly all variants. Both can prioritize all SNVs, and CADD can prioritize smaller indels, while VVP can prioritize all indels that can be annotated by VEP. VVP is built on the VAAST likelihood and utilizes lookups based on healthy human variation to prioritize variants. I have shown that not only is VVP able to prioritize more variants than CADD, it is faster and more accurate. Thus, VVP is the leading tool for human variant prioritization.

VVP scales well to large datasets because of the organization of the lookups and because the computational work required to process a single variant is unchanged with respect to the number of individuals in the background and very nearly unchanged with respect to the number of individuals in the target. A very exciting future direction is to develop a burden test using the VVP framework. This would provide a scalable solution to performing burden tests with cohorts that have tens of thousands of cases and controls.

Metagenomics

Metagenomics holds enormous promise to revolutionize our understanding of the microbial world and pathogen diagnostics by providing a hypothesis free method to query microorganisms in an environmental sample (Brady & Salzberg, 2009). Of particular importance is using metagenomics to find microorganisms that are responsible for human illness from a fluid or tissue sample.

Modern metagenomics produce datasets with tens of millions of short reads from an environmental sample. From a computational perspective, the metagenomics problem is to classify every read with as much taxonomic precision as possible. BLAST contains the functionality necessary to classify reads; however, it is too slow to be practical on large read sets that are now common. Faster approaches are necessary (Wood & Salzberg, 2014).

I created Taxonomer: a collection of tools that enable rapid analysis of metagenomics datasets. Taxonomer provides functionality to classify reads in both nucleotide and protein space and provides RPKM estimates of host gene expression. A website using the iobio framework provides easy and rapid access to Taxonomer's capabilities. Extensive benchmarking has shown that Taxonomer is not only more comprehensive in its classification abilities than any other single tool, but is also extremely fast and provides accurate results.

Central to Taxonomer's speed and accuracy is a novel k-mer-based weighting scheme that provides a rapid and powerful way to classify read sequences. In addition, a novel transformation enables the same algorithms that classify reads in nucleotide space to classify the same reads in protein space with only a moderate penalty in memory usage and an extremely small time penalty. Because of the powerful mapping capability of the k-mer-based weighting scheme, Taxonomer is also able to rapidly quantify gene expression with accuracy equal to that of the best available transcript profiling software. Taxonomer's extensive capabilities make it a tool that is able to work effectively in answering many different questions important in the application of metagenomics to both research and medical diagnostics.

Summary and future directions

In my dissertation, I have presented effective computational approaches and applications to a wide variety of data analysis problems in the biological sciences. Specifically, I have presented compelling solutions to image analysis, human variant prioritization, and metagenomics. All the methods and applications I have presented in this dissertation have exciting future possibilities, in particular in the areas of human variant prioritization and metagenomics. Extending VVP to include a burden test would provide a highly scalable solution to identify genes responsible for disease in settings with extremely large numbers of target and background individuals. Taxonomer can be further improved with better sequence databases to improve classification accuracy and making the web interface as comprehensive as possible in its analysis capabilities while keeping it relatively simple to use.

References

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–6. doi:10.1038/nmeth.1358

Flygare, S., Campbell, M., Ross, R. M., Moore, B., & Yandell, M. (2013). ImagePlane: An automated image analysis pipeline for high-throughput screens using the planarian schmidtea mediterranea. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *20*(8), 583–92. doi:10.1089/cmb.2013.0025

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations : Biological impact and theoretical interpretation. *Protein Science*, *23*, 1650–1666. doi:10.1002/pro.2552

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–5. doi:10.1038/ng.2892

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. doi:10.1186/gb-2014-15-3-r46