Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes

Michael F. Lin* Ameya N. Deoras† Matthew D. Rasmussen† Manolis Kellis*†

Abstract

Comparative genomics on multiple related species is a powerful methodology for the discovery of functional genomic elements, and its power should increase with the number of species compared. We use twelve *Drosophila* genomes to study the power of both single-sequence and comparative genomics metrics to distinguish between protein-coding and non-coding regions. We find that species at a broad range of evolutionary distances are comparably effective informants for pairwise comparative gene identification, and that the pairwise methods we studied achieve higher specificity but lower sensitivity than advanced single-sequence metrics. We find that multi-species analysis leads to higher discriminatory power, which progressively increases as additional informant species are used. Overall, multi-species metrics robustly outperform single-sequence metrics, especially on shorter exons ($\leq$240nt), which are common in animal genomes. Lastly, we find that single-sequence and comparative metrics capture largely independent features of protein-coding genes, and that combining them leads to hybrid metrics with greater discriminatory power. Our results have implications for comparative genomics analyses in any species, including the human.

Introduction

Computational methods for *de novo* gene identification play a major role in understanding the protein-coding gene catalog of any species. In higher eukaryotes, gene predictors use probabilistic models such as generalized hidden Markov models (GHMMs) to predict the exon-intron gene structures prevalent in these genomes, based on compositional properties and sequence signals characteristic of protein-coding exons (reviewed in Zhang, 2002). These systems have also been extended to incorporate comparative genomics signals, taking advantage of the preferential conservation of genes to greatly increase their accuracy (reviewed in Brent, 2005). However, because any type of evidence used in a GHMM must be explicitly modeled in a joint probability distribution over the hidden state and output sequences, they are limited in the types of features they

*Broad Institute of MIT and Harvard. 7 Cambridge Center, Cambridge, MA 02139

†Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology. 32 Vassar St., Cambridge, MA 02139

can incorporate while still allowing tractable inference algorithms (Sutton and McCallum, 2006). This imposes constraints on the form and character of genomic signals used in GHMM-based gene predictors, perhaps limiting their overall discovery power.

Recent advances in machine learning have led to a new family of probabilistic sequence labeling models that have far greater flexibility than GHMMs to incorporate richer discriminative features. A *semi-Markov conditional random field* (SMCRF) can, in principle, directly incorporate any metric that produces a score for any interval in the sequence under analysis (Lafferty et al., 2001; Sarawagi and Cohen, 2005). Initial efforts to build *de novo* gene structure predictors using SMCRFs have already yielded promising results (Bernal et al., 2007; Decaprio et al., 2007; Gross et al., 2007; Vinson et al., 2007). These studies have focused primarily on algorithms for training SMCRFs and their advantages over GHMMs, and less on exploring the metrics they use to discriminate between coding and non-coding regions.

In this paper, we study discriminative metrics for gene identification that can be incorporated into gene identification systems. Given a region of the genome and, when available, its alignment across related species, these metrics produce a score that indicates how likely it is to be protein-coding. We study the ability of single-sequence, pairwise, and multi-species methods to discriminate between coding and non-coding regions, and seek to determine which methods and methodological choices lead to the greatest discovery power. For comparative approaches, we also evaluate the impact of choices that precede the application of any individual metric, such as the genome alignment strategy, and the evolutionary distance and number of "informant" species used to analyze the "target" genome.

To simplify the exposition, we apply these metrics to a binary classification problem, in which we are given a series of genomic regions and we must decide whether each one is protein-coding or non-coding. This allows us to study the power of each metric outside of the "black box" of GHMM and SMCRF training and decoding algorithms, which admit little intuition. For example, while it stands to reason that multiple species should be better than pairwise comparisons in gene predictors, initial efforts to this end have been met with limited success (Gross and Brent, 2006). It remains unclear whether this is due to fundamental issues relating to the choice and alignment of

informant species, or whether it merely reflects limitations of the algorithms used thus far (Brent, 2005). By separating and carefully studying the discriminative metrics we build into such models, we aim to establish basic expectations and provide guidance for improving their performance.

We also note that discriminative metrics that reliably distinguish protein-coding sequences are useful for several purposes other than *de novo* gene prediction. For example, they can be used to determine whether experimentally derived transcript sequences are likely to be protein-coding mRNAs (Frith et al., 2006a; Liu et al., 2006). Additionally, they can be used to evaluate and refine existing annotations. We have previously used discriminative metrics to substantially revise the *Saccharomyces cerevisiae* gene annotations (Kellis et al., 2003), and in a related study, we use metrics presented in this paper to propose revisions to over 10% of the established gene annotations for the *Drosophila melanogaster* genome (Lin, Carlson, Crosby *et al.*, submitted).

In the next section, we describe each of the discriminative metrics we chose to study. We then describe a large dataset we constructed to benchmark these metrics, consisting of tens of thousands of coding and non-coding sequences aligned across twelve *Drosophila* genomes. We measure the discriminatory power of each metric and how it varies with sequence length, phylogenetic distance, total number of informant sequences, and the genome alignment strategy. We also investigate the redundancy and independence of the metrics, and show that they can be combined to substantially increase total discriminatory power. Finally, we discuss the overall strategic implications of our results for comparative approaches to gene identification.

Discriminative metrics for gene identification

We evaluate several well-known methods for gene identification, and also introduce and evaluate several metrics that we have developed. We outline these here, and provide full implementation details in supplemental methods.

Pairwise comparative metrics

Most initial efforts at comparative gene identification used a single informant genome to support the annotation of a target genome (Alexandersson et al., 2003; Badger and Olsen, 1999; Batzoglou

et al., 2000; Korf et al., 2001; Meyer and Durbin, 2002; Mignone et al., 2003; Parra et al., 2003; Zhang et al., 2003). These systems typically do not provide a modularized discriminative metric that can be applied in isolation, so we selected several metrics that we believe capture the essential properties of coding sequence evolution that they observe:

- K_A/K_S is the ratio of the rate of non-synonymous substitutions per non-synonymous site (K_A) to the rate of synonymous substitutions per synonymous site (K_S) in a pairwise alignment. Genes and exons under purifying selection tend to have $K_A/K_S \ll 1$, indicating that synonymous substitutions are far more common than non-synonymous substitutions. We used the method of Nei and Gojobori (1986) to estimate this value, which is simple and widely used although it is known to have certain inherent biases (Yang and Nielsen, 2000). (see also dN/dS test, below)

- **Codon Substitution Frequencies (CSF)** is a metric we developed which, like K_A/K_S, observes biases in codon substitutions in alignments of coding sequences, rewarding synonymous substitutions and penalizing non-synonymous substitutions. CSF assigns a score to each codon substitution in a pairwise alignment based on quantitative estimates of the frequencies at which every pair of codons is substituted between the two species being analyzed, leading to finer distinctions between different codon substitutions. For example, non-synonymous substitutions that preserve amino acid chemical properties are penalized less than disruptive non-synonymous substitutions, and nonsense substitutions are very strongly penalized.

- **Reading Frame Conservation (RFC)**, which we proposed for use in yeast comparative genomics (Kellis et al., 2003), observes the strong bias of indels within coding regions to be multiples of three in length, in order to preserve the codon reading frame of translation. The RFC score is computed as the percentage of nucleotides in the informant sequence that occur in the same reading frame offset as the target sequence (Kellis et al., 2004).

- **TBLASTX**, which identifies protein sequence similarity (Altschul et al., 1997), is widely used as supporting evidence for gene identification (Crollius et al., 2000; Parra et al., 2003; Zhang et al., 2003). We ran TBLASTX for each of the sequences in our dataset against the

repeat-masked genome assembly of the informant species, and took the "bit score" of the best hit as the score for each sequence. (Unlike the other pairwise metrics, TBLASTX does not require existing alignments of the informant sequences.)

- **Sequence conservation**. Finally, we implemented a baseline metric that simply measures the percent identity between the target and informant sequences, to account for alignability and conservation as indicators of function. Since animal genomes contain many more conserved sequences than can be explained by protein-coding exons (Mouse Genome Sequencing Consortium, 2002; Siepel et al., 2005), we did not expect this metric to achieve high accuracy.

Multi-species comparative metrics

Several gene identification systems capable of integrating multiple species have been developed (Gross and Brent, 2006; Pedersen and Hein, 2003; Siepel and Haussler, 2004). Although the use of multiple informant species should in principle provide more evidence than any single informant (Eddy, 2005; Margulies et al., 2007; Thomas et al., 2003), investigations of whether gene predictor performance indeed improves with multiple species in practice have yielded mixed results (Brent, 2005; Gross and Brent, 2006; Pedersen and Hein, 2003). We selected the following metrics that use multiple alignments, again attempting to capture the essential properties of coding sequence evolution that gene predictors observe:

- **dN/dS test**. The dN/dS test uses a probabilistic phylogenetic algorithm to estimate the rate of non-synonymous substitutions (dN) relative to the rate of synonymous substitutions (dS) in a multiple sequence alignment (Nekrutenko et al., 2002; Yang and Bielawski, 2000). Like the K_A/K_S ratio, genes and exons under purifying selection tend to have $dN/dS \ll 1$, although the exact inferred value of dN/dS plays a secondary role in this test. Instead, it is a statistical test of the hypothesis that $dN/dS < 1$, carried out by comparing the likelihoods computed by a sophisticated model of codon evolution fit with dN/dS fixed at one to a corresponding model in which dN/dS is estimated by maximum likelihood. These models are implemented in PAML (Yang, 1997).

- **Codon Substitution Frequencies (CSF)**. In order to avoid the high computational expense of maximum likelihood-based phylogenetic algorithms, we developed an *ad hoc* strategy to extend the CSF metric to incorporate evidence from multiple species, by taking the median of the scores of the pairwise substitutions in each codon column of the multiple alignment, and then summing these medians across the sequence (see supplemental methods for details).

- **Reading Frame Conservation (RFC)**. The RFC test uses a simple voting scheme to combine evidence from multiple informant species (Kellis et al., 2004, 2003). Each informant species casts a vote of $+1$ if its pairwise RFC score is above a certain cutoff, -1 if it is below, or 0 if there is insufficient alignable sequence. The cutoff is chosen for each species by examining the distribution of the RFC score in known genes and non-coding regions. The multiple-informant RFC score takes on integer values in $[-n, n]$, where n is the number of informant species.

- **Sequence conservation**. As with the pairwise metrics, we implemented a simple sequence conservation metric as a baseline for comparison. We assigned a score to each nucleotide column of a multiple alignment corresponding to the largest fraction of species having the same nucleotide in that column (plurality). We then took the average of these scores for each nucleotide column in the alignment as the score for each sequence.

Single-sequence metrics

We also included several single-sequence metrics in our benchmarks, to gauge their performance relative to comparative methods. These are based on compositional biases and periodicities that distinguish protein-coding from non-coding sequences, which form the foundation of most *de novo* gene predictors (Zhang, 2002). Previous studies have benchmarked many single-sequence metrics with various parameter settings (Gao and Zhang, 2004; Saeys et al., 2007), and we chose only a representative set here.

- **The Fourier transform** measures the strength of the three-base periodicity in coding sequences, which is a result of biases in the genetic code favoring certain nucleotides in certain

positions (Anastassiou, 2001).

- **Codon bias** observes the unequal usage of synonymous codons in protein-coding sequences, which results in part from how different synonymous codons affect translation efficiency (Akashi, 2001).

- **Interpolated context models (ICMs)** are generative probabilistic models used to observe reading frame-dependent biases in the frequencies of k-mers in coding sequences, simultaneously for several different k-mer sizes (Delcher et al., 1999).

- **Z curve** also observes reading frame-dependent biases in k-mer frequencies, but uses a discriminative approach based on Fisher linear discriminant analysis (Gao and Zhang, 2004).

Experimental design: genome-wide benchmarks for discriminative metrics

In order to benchmark the discriminatory power of each of the metrics we studied, we developed a test set consisting of exons and non-coding regions in the genome of the fruit fly *Drosophila melanogaster*. We chose to study the fly genome for several reasons. First, the fly gene annotations are of very high quality, following a century of classical genetics, large-scale transcript sequencing projects (Rubin et al., 2000; Stapleton et al., 2002a,b), and extensive manual curation (Misra et al., 2002). Second, the fly exhibits many of the biological complexities that challenge gene identification strategies in vertebrates, such as frequent lengthy introns, abundant conserved non-coding elements, genes nested within introns of other genes, and unusual gene structures such as polycistronic transcripts. Third, the recent sequencing of ten *Drosophila* genomes (*Drosophila* Comparative Genome Sequencing and Analysis Consortium, submitted), in addition to *D. melanogaster* (Adams et al., 2000) and *D. pseudoobscura* (Richards et al., 2005), provides a rich set of informants with which to evaluate genome-wide comparative genomics methods (Figure 1).

We selected 2,734 genes from FlyBase annotation Release 4.3 (Crosby et al., 2007), containing 10,722 non-overlapping coding exons. We also selected 39,181 random non-coding regions, with

the same length and strand distribution as the exons (see supplemental methods for details). We chose this strategy of randomly sampling genes and selecting all exons of those genes, rather than directly sampling exons, to facilitate studying how the power of each metric varies across different functional categories of genes. Although not by design, the length distribution of sequences in our test set (median = 224nt, mean = 404nt, sd = 570nt) is very similar to the length distribution of exons in the genome (median = 220nt, mean = 408nt, sd = 568nt).

We then extracted each of the regions in our dataset from whole-genome sequence alignments of the twelve fly genomes. We used two different sets of genome alignments. The first was generated by MULTIZ (Blanchette et al., 2004), which uses local alignments of high-similarity regions; the second was generated by the Mercator orthology mapper (C. Dewey and L. Pachter) and MAVID sequence aligner (Bray and Pachter, 2004), and is based on the identification of orthologous segments in each genome by conserved gene order (synteny). We used both sets of alignments to compare the effectiveness of these strategies for comparative gene identification.

For each metric, we scored all the 49,903 regions in our test set (10,722 exons and 39,181 non-coding regions) and then measured its ability to correctly classify them as coding or non-coding. We used four-fold cross-validation to train and apply the metrics that require training data. We evaluated the performance of each metric by examining receiver-operator characteristic (ROC) curves showing its sensitivity and specificity[1] at different score cutoffs. Based on the ROC curve for each metric, we also computed two different summary error values, capturing different aspects of their overall classification performance:

- The *minimum average error* (MAE) is the average of the false negative rate (1-sensitivity) and the false positive rate (1-specificity), at the cutoff where this average is minimized; intuitively, this is the "elbow" of the ROC curve.

- The *area above the curve* (AAC) is the area lying above the ROC curve in the unit square.

[1]Here and throughout this paper, we use the term *specificity* as it is defined in binary classification problems: the fraction of true negatives that are correctly classified as negative. This differs from the common usage of the term in the gene prediction field to refer to the fraction of the examples classified as positive that are true positives (which we would call *precision*). Since the fly genome is about 20% protein-coding, the precision can be roughly estimated as $\frac{Sn}{Sn+4\cdot(1-Sp)}$ although this ignores various issues of segmentation and strandedness that would need to be addressed in order to use these metrics to make exon predictions. Additionally, we use the term *false positive rate* to mean 1-Specificity, or the fraction of true negatives incorrectly classified as positive.

The MAE has an intuitive interpretation as the fraction of examples that are incorrectly classified (if the positive and negative classes are the same size), but it represents only a single point on the ROC curve. The AAC summarizes more information about classification performance over all sensitivity/specificity regimes, but it lacks a simple interpretation. For both error measures, a perfect classifier would have zero error, and a random classifier would produce 0.5 error. While these statistics cannot represent all the sensitivity/specificity tradeoffs captured by the ROC curve, they provide a simple means of comparing the performance of different metrics and methodological choices.

Results

Overall discovery power of each metric

We first compared the overall performance of the metrics (Figure 2). All of the metrics we evaluated demonstrated high classification performance, but some general trends were apparent. The comparative metrics (using the MULTIZ alignments of all twelve fly genomes) generally outperformed the single-sequence metrics (except for the baseline sequence conservation metric). For example, the best comparative metric resulted in 24% lower error than the best single-sequence metric (0.050 MAE for the dN/dS test vs. 0.065 for Z curve). Different metrics were preferable at different sensitivity/specificity tradeoffs. For example, the CSF and dN/dS metrics achieved the highest specificity (99.9% for CSF) even at fairly high sensitivities (85.2%). RFC tended towards higher sensitivity and lower specificity than CSF and dN/dS.

We also compared the pairwise metrics, using the best pairwise informant (*D. ananassae*; we investigate different pairwise informants below), and found similar trends (Supplemental Figure 1). For example, CSF and K_A/K_S performed comparably, showing the highest specificity, while RFC tended towards higher sensitivity and lower specificity. TBLASTX underperformed K_A/K_S, CSF, and RFC, but it was better than our baseline conservation metric. Notably, none of the pairwise comparative metrics outperformed the best single-sequence metric (Z curve) according to MAE and AAC error (Figure 3b), and they exhibited generally lower sensitivity. CSF and K_A/K_S were,

however, able to achieve higher specificity at a moderate sensitivity tradeoff. For example, at 80% sensitivity, CSF had a nearly ten-fold lower false positive rate than Z curve (0.15% and 1.39%); the specificity of CSF exceeded Z curve at less than 85% sensitivity, compared to 93% sensitivity at Z curve's MAE point.

Comparative methods are strongly preferred for short exons

We next assessed each metric's discriminatory power for different sequence length categories (Figure 2c). All of the metrics performed better on longer sequences than shorter sequences. Single-sequence metrics performed comparably or slightly better than comparative methods for long sequences (>240nt), but comparative methods strongly outperformed single-sequence metrics on shorter sequences. For example, in the length range of 181-240nt (which includes the median exon length) the best comparative metric resulted in 51% lower error than the best single-sequence metric (0.027 MAE for the dN/dS test and 0.056 MAE for Z curve). In the shorter length range of 121-180nt, the best comparative metric resulted in 60% lower error than the best single-sequence metric (0.029 MAE for CSF and 0.073 MAE for Z curve). Different comparative methods were also preferred at different lengths. For example, CSF strongly outperformed the dN/dS test on the shortest sequences ($\leq$60nt), while they performed comparably on longer sequences.

Comparing alignment strategies

We then compared the above results, based on MULTIZ local similarity-based alignments, with the corresponding results based on the synteny-anchored alignments generated by Mercator/MAVID. Overall, the Mercator/MAVID alignments led to similar *trends* in the performance of the metrics relative to each other and across different sequence lengths. However, the two alignments often led to different *absolute* levels of performance.

We expected the local alignment approach to give higher sensitivity than the synteny-anchored alignments, since it should be better able to align exons that have undergone rearrangements (Blanchette, 2007). Indeed, we found that the MULTIZ alignments led to higher sensitivity than the Mercator/MAVID alignments (e.g. 90% vs. 87% for CSF at 99% specificity). Conversely, we

expected the synteny-anchoring approach used by Mercator/MAVID to give higher specificity than the local alignment approach of MULTIZ, since it may generate fewer spurious non-orthologous alignments (Blanchette, 2007). However, we found that while the Mercator/MAVID alignment could lead to slightly higher specificity, it did so only at disproportionate sensitivity tradeoffs. For example, with the baseline sequence conservation metric, specificity using the Mercator/MAVID alignments exceeded that of the MULTIZ alignments only at lower than 58% sensitivity (compared to 80% sensitivity at the MULTIZ-based MAE point). Similarly, with RFC, specificity resulting from the Mercator/MAVID alignments was greater only at lower than 63% sensitivity (compared to 92% MAE sensitivity).

Overall, the considerably lower sensitivity resulting from the Mercator/MAVID alignments was reflected in uniformly worse MAE and AAC error statistics (Supplemental Figure 2). Moreover, for sensitivity ranges above $\sim 60\%$, they also led to lower specificity than the MULTIZ alignments. We therefore focused on the MULTIZ alignments for the remainder of our analysis, although we note that these empirical observations may be highly dependent on parameter settings of the genome alignment programs, and further investigation of the tradeoffs of their strategies is needed.

A wide range of phylogenetic distances are effective in pairwise analysis

To investigate which species are the most and least effective informants for gene identification, we evaluated each pairwise comparative metric using informant genomes at increasing evolutionary distance from *D. melanogaster*. We applied each metric to pairwise alignments of *D. melanogaster* with *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, and *D. grimshawi*, each representing various clades within the genus *Drosophila* (Figure 1).

We found that, at least according to MAE and AAC error values, *D. ananassae* was overall the most effective informant, outperforming other species on most metrics. However, inspection of the corresponding ROC curves often revealed a more complex situation, with multiple species showing similar performance, and sometimes higher for certain sensitivity/specificity tradeoffs. For example, with K_A/K_S, *D. ananassae* and *D. willistoni* performed comparably, with *D. ananassae* leading to slightly higher sensitivity and *D. willistoni* leading to slightly higher specificity (Figure

3a). Similarly, with RFC, closely related species led to slightly higher sensitivities, and more distant species led to slightly higher specificities (Supplemental Figure 3). Hence, while *D. ananassae* was overall the most effective informant, it did not *robustly* outperform the other pairwise informants we studied. The only exception was *D. erecta*, the most closely related to *D. melanogaster* of the species we studied. *D. erecta* was consistently less informative than the others, leading to the lowest classification performance on most of the pairwise metrics (although it still performed well in absolute terms, e.g. 10% MAE using CSF).

To investigate more distant species for which we lacked whole-genome alignments, we also applied TBLASTX to the genomes of the mosquito (Holt et al., 2002) and honeybee (Honeybee Genome Sequencing Consortium, 2006). We found that these species led to much worse performance than the *Drosophila* species as informants for *D. melanogaster* (Figure 3b).

We conclude that a broad range of species within the genus *Drosophila* (outside of the *melanogaster* subgroup) make effective pairwise informants for gene identification in *D. melanogaster*, while the mosquito and honeybee, the next most closely related species with fully sequenced genomes, are likely to be too distant for this application. These findings are consistent with a previous smaller-scale study of comparative gene identification power in flies (Bergman et al., 2002), and previous theoretical and simulation studies suggesting that, while some mathematically optimal distance may exist, species at a broad range of phylogenetic distances should be comparably effective informants for identifying exons and other conserved elements (Eddy, 2005; Zhang et al., 2003).

Multi-species comparisons lead to higher performance

We next investigated the effectiveness of increasing numbers of informant species on the metrics that can use multiple informants. We evaluated each metric using subsets of the available species corresponding to increasingly broad clades within the genus *Drosophila* (see phylogeny in Figure 1): the *melanogaster* subgroup (5 species including *D. melanogaster*), the *melanogaster* group (6 species), the *melanogaster* and *obscura* groups (8 species), the subgenus *Sophophora* (9 species), and finally all 12 species of the genus *Drosophila*.

We found that for each of the metrics we benchmarked in this way, discriminatory power tended

to increase as additional informant species were used (Figure 4a). In contrast to our previous pairwise analysis, in which the most distant *Drosophila* informants led to similar or slightly worse performance than closer species, *adding* informants at increasing distances led to a clear trend in higher classification performance. The dN/dS test, RFC, and the sequence conservation metric each showed a smooth progression of increasing performance with each successively larger group of informant species. For example, starting from the four informants within the *melanogaster* subgroup, the dN/dS test achieved an MAE of 0.103. With the addition of each successive group of informants, the MAE was reduced relatively by 35%, 43%, 48%, and finally by 52%. CSF showed a similar trend through the subgenus *Sophophora*, but did not appear to benefit from the subsequent addition of the final three informants of subgenus *Drosophila*. This saturation effect may be related to the use of the median column score in its implementation, and suggests that this aspect can be improved upon for larger numbers of informants.

With a sufficient number of informants, the multi-species metrics surpassed single-sequence metrics according to MAE (Figure 4b). This also stands in contrast to our pairwise analysis, in which no informant enabled any comparative metric to outperform the best single-sequence metric (Z curve). CSF exceeded the performance of Z curve once we used at least six species ($\geq$ 1.3 sub/site), dN/dS with at least eight species ($\geq$ 1.9 sub/site), and RFC, using its simplistic vote-tallying scheme, with all twelve species (4.1 sub/site). The baseline sequence conservation metric never outperformed Z curve, although its performance also increased with additional species. (We note that while these results show that a certain number of informants is *sufficient*, they do not imply that they are all *necessary* to achieve some level of performance; removing informants that contribute very little independent branch length might not substantially reduce performance.)

In most cases, the four informants of the *melanogaster* subgroup together yielded worse performance than pairwise analysis with the best pairwise informant, *D. ananassae*. In contrast, all of the informant clades that combined *D. ananassae* with more distant species led to better performance than any pairwise analysis. This affirms our earlier conclusion, based on a pairwise analysis with *D. erecta*, that the species within the *melanogaster* subgroup are sub-optimal informants for the metrics we studied, presumably because they are too closely related to *D. melanogaster*. Indeed,

the neutral distance of *D. ananassae* from *D. melanogaster* is 1.0 substitutions per neutral site, while the *total* independent branch length provided by the four *melanogaster* subgroup informants is only 0.4 sub/site.

Characterizing genes that comparative methods fail to detect

It is well-known that genes in certain categories of biological function tend to be faster-evolving than average (Holt et al., 2002; Honeybee Genome Sequencing Consortium, 2006; Richards et al., 2005; Zdobnov et al., 2002). We set out to investigate whether comparative metrics are therefore unable to distinguish such genes from non-coding regions. We obtained Gene Ontology (GO) annotations (Ashburner et al., 2000; Misra et al., 2002) for each of the 2,734 genes comprising our test set. For each of the 192 GO terms represented by at least thirty genes in our test set, we determined the fraction of those genes with at least one exon scoring above a stringent cutoff ("detected genes").

We found that all of the functional categories we investigated had very high detection rates (Supplemental Table 1). For example, with a CSF cutoff corresponding to 85% exon sensitivity and 99.9% specificity using all twelve fly genomes, the overall fraction of detected genes was 92%, and the detection rates surpassed 90% for all but two functional categories: serine-type endopeptidase activity (89% detected genes) and its superset, serine-type peptidase activity (86%). Serine proteases play key roles in insect innate immunity, and some likely evolve under positive selection (Holt et al., 2002; Jiggins and Kim, 2007; Reichhart, 2005). Several other categories that intuition suggests might relate to more rapidly evolving genes, however, were not problematic, including immune response (94%), gametogenesis (95%) and G-protein coupled receptor activity (100%).

Instead, comparative metrics had the most difficulty detecting genes of unknown function. Three GO terms indicating unknown function (unknown cellular component, molecular function, and biological process) had only 67%, 61%, and 60% detected genes. In fact, of the genes that were not detected at this cutoff, 85% were of unknown function or lacked any GO term, compared to 49% of all the genes in our dataset. These trends held, qualitatively, for all of the comparative metrics and cutoffs we investigated (Supplemental Table 1).

Overall, these results indicate that comparative methods using the twelve fly genomes were

able to detect the vast majority of genes in all of the functional categories we investigated (which were represented by at least 30 genes in our dataset; a larger sample might reveal more specific functional categories that are, in fact, very difficult for comparative methods to detect). They had much greater difficulty detecting genes of unknown function, which may be under less selective constraint overall (*Drosophila* Comparative Genome Sequencing and Analysis Consortium, submitted; Bergman et al., 2002) but could also include a higher proportion of incorrect or spurious annotations (Lin, Carlson, Crosby *et al.*, submitted). Interestingly, Z curve, a single-sequence metric, also showed much lower sensitivity to genes of unknown function (Supplemental Table 1), suggesting that these genes, if they are correctly annotated, tend to be unusual in several ways.

Independence and combinations of the metrics

While each of the metrics we studied exhibited unique performance characteristics, some measure similar fundamental lines of evidence, and thus may tend to err on the same examples. For example, K_A/K_S, the dN/dS test, and CSF all observe the distinctive biases in codon substitutions in protein-coding sequences, and all may tend to miss genes with low peptide identity in the informant species. In contrast, RFC observes patterns of insertions and deletions that are essentially orthogonal to codon substitutions. Additionally, the single-sequence metrics observe compositional biases and periodicities that are ignored by the comparative metrics.

We investigated the independence of the metrics, indicated by how differently they rank the exons in our test set, using a dimensionality reduction technique called multidimensional scaling (MDS; Cox and Cox, 2001). This analysis led to a two-dimensional visualization shown in Figure 5b, in which each point represents one of the metrics and the distance between the points approximately represents their dissimilarity. As expected, we found that the dN/dS test and CSF behaved very similarly, while RFC was clearly distinct. The sequence conservation metric was separate from each of these, while TBLASTX clustered with CSF and dN/dS. The four single-sequence metrics formed two additional clusters distinct from the comparative metrics, suggesting that they capture largely independent properties of protein-coding genes.

The relative independence of several of the metrics suggests that combining them could lead to

higher performance. We selected five metrics representing each of the MDS clusters (CSF, RFC, sequence conservation, Z curve, and codon bias) and combined them using cross-validated linear discriminant analysis (LDA), which produces a composite score for each sequence using a linear combination of the individual metrics. As expected, the hybrid metric outperformed any of its inputs: by MAE error, the LDA hybrid resulted in 27% lower error than its best input metric (0.040 MAE for LDA vs. 0.055 for CSF). The hybrid metric demonstrated much higher sensitivity than any of its input metrics (Figure 5a), and higher specificity than all of the input metrics except CSF. We obtained almost identical results using a second hybrid metric based on a linear support vector machine instead of LDA. Thus, although CSF and the dN/dS test remain the methods of choice for the highest specificity, the hybrid metrics achieved higher overall performance. Additionally, their position in the MDS visualization (Figure 5b) suggests that they do so by combining the distinct information from the individual metrics.

Discussion

In this paper, we investigated discriminative metrics for distinguishing protein-coding sequences from non-coding sequences. We found that multi-species comparative methods outperform single-sequence metrics, particularly on short sequences ($\leq$ 240nt). On the other hand, the pairwise comparative methods we studied achieved higher specificity, but did not outperform advanced single-sequence metrics overall. We found that a broad range of species within the genus *Drosophila* are comparably effective pairwise informants for *D. melanogaster*, in agreement with theoretical predictions. We showed that adding more species to comparative analysis progressively increased discovery power for a variety of different methods. Finally, we showed that several comparative and single-sequence metrics can be combined into a hybrid metric more powerful than any of its parts.

Among the three multi-species comparative metrics we studied (CSF, the dN/dS test, and RFC; excluding the baseline sequence conservation metric), none strictly outperformed the others. RFC tended towards lower specificity but higher sensitivity than CSF and the dN/dS test. CSF was more effective than the dN/dS test on the shortest exons, but they performed comparably overall,

and both achieved near-perfect specificity at moderate sensitivity tradeoffs. We developed CSF as a simpler alternative to the computationally expensive phylogenetic algorithms upon which the dN/dS test is based, and we consider it successful in this respect, considering their comparable results and their total compute times (on our dataset, several minutes for CSF vs. a few weeks for the dN/dS test using PAML). On the other hand, our tests with different numbers of informant species suggest that the CSF method may benefit from future improvement in order to take advantage of ever-larger numbers of informants. It is also likely that RFC, which uses a simplistic vote-tallying scheme to combine evidence from multiple species, can be significantly improved upon. The fact that these relatively simple methods outperformed advanced single-sequence metrics, and even competed with maximum-likelihood phylogenetic algorithms, speaks to the power of the underlying comparative data.

Selection of informants for comparative gene identification

We found that species ranging from 1.0–1.4 substitutions per neutral site from *D. melanogaster* are comparably effective informants for pairwise gene identification, with slight preference given to the closer end of this range. This "optimal" range might extend both for closer species (between *D. erecta* and *D. ananassae*) and more distant species (between *D. grimshawi* and *A. gambiae*), but these distances were not explored in the currently sequenced species. This range is comparable to the distance from human of the opossum (0.8 sub/site), chicken (1.1 sub/site), and lizard (1.3 sub/site), suggesting that species more distant than the eutherian mammals, the farthest of which are less than 0.5 sub/site (Figure 1), may prove to be excellent informants for human gene identification.

Clearly, however, the neutral substitution rate is not the only consideration that should guide comparative informant selection. The reliable alignment of genomes at the "optimal" distances we identified may be very difficult without relatively expensive, draft-quality or better informant genome assemblies (Margulies et al., 2005), which we were fortunate to have available for *Drosophila*. Accurate alignment is further complicated by the greater degree of chromosomal rearrangement and gene family evolution that may be expected at such distances (*Drosophila* Comparative Genome

Sequencing and Analysis Consortium, submitted). More fundamentally, distant species share less in common biologically; indeed, the 12 *Drosophila* species were selected in part to represent the diverse ecological niches they occupy (Markow and O'Grady, in press) and the neutral distance they span (approximately corresponding to the distance between human and reptiles).

Thus, while our results suggest that such distant species may nonetheless be highly informative given high-quality sequences and alignments, future empirical studies should investigate the use of many species at closer distances, such as those represented by the eutherian mammals, for gene identification. Our results using different subsets of informants clearly show that more informants lead to more discovery power, but they do not fully address the question of how many informants at eutherian mammal-like distances would be needed to exceed the performance of the more distant pairwise informants. Additionally, we note that a separate category of "phylogenetic shadowing" techniques (Boffelli et al., 2003), which are designed to operate at even closer distances (e.g. within primates), may be helpful for detecting some of the most rapidly evolving or recently evolved genes.

Implications for gene prediction strategies

An important application of the metrics we have studied in this paper will be their integration into *de novo* gene structure predictors based on semi-Markov conditional random fields, which can combine multiple discriminative metrics in a manner not unlike our LDA hybrid. Our results suggest that these systems should be able to use multiple informant species and multiple metrics to identify protein-coding sequences with higher accuracy, especially on short exons. Still, it is not obvious that these trends in the metrics' performance will necessarily translate into higher-accuracy prediction of complete gene structures, since the latter also strongly depends on the detection of splice sites and other sequence signals (Zhang, 2002). More fundamentally, the probabilistic models used in gene predictors make simplifying assumptions about gene structures that lead to many incorrect predictions, and that cannot be relaxed just by using more powerful metrics. For example, they currently cannot predict nested and interleaved genes, which are increasingly understood to be fairly common in higher eukaryotic genomes (ENCODE Project Consortium, 2007; Karlin et al., 2002; Misra et al., 2002; Yu et al., 2005), since these structures violate Markov independence assumptions.

A similar challenge is presented by alternative splice isoforms with mutually exclusive exons that do not splice to each other in-frame.

In addition to *de novo* gene structure prediction, the methods we have studied can serve other important applications, such as evaluating and refining existing annotations, and searching the genome for coding regions that are systematically missed or erroneously modeled by other methods. For example, the effectiveness of comparative methods for detecting short coding regions may prove crucial in identifying short proteins, several families of which are known to serve important biological roles, but which have been systematically under-represented in genome annotations (Frith et al., 2006b; Galindo et al., 2007). They also provide a promising way to search for gene structures that violate traditional assumptions entirely, such as stop codon readthrough, translational frameshifts and polycistronic transcripts (Lin, Carlson, Crosby *et al.*, submitted).

Acknowledgements

We thank Huy L. Nguyen for informatics assistance; David DeCaprio, Jade Vinson and James Galagan for helpful discussions regarding SMCRFs; and Alexander Stark and Pouya Kheradpour for additional comments and discussions.

References

Adams, M., Celniker, S., Holt, R., Evans, C., Gocayne, J., Amanatides, P., Scherer, S., Li, P., Hoskins, R., Galle, R., *et al.*, 2000. The genome sequence of drosophila melanogaster. *Science*, **287**(5461):2185.

Akashi, H., 2001. Gene expression and molecular evolution. *Curr Opin Genet Dev*, **11**(6):660–666.

Alexandersson, M., Cawley, S., and Pachter, L., 2003. Slam: Cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Res.*, **13**(3):496–502.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17):3389–3402.

Anastassiou, D., 2001. Genomic signal processing. *IEEE Signal Processing Magazine*, **18**:8–20.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.*, 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1):25–29.

Badger, J. H. and Olsen, G. J., 1999. Critica: coding region identification tool invoking comparative analysis. *Mol Biol Evol*, **16**(4):512–524.

Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B., and Lander, E. S., 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.*, **10**(7):950–958.

Bergman, C., Pfeiffer, B., Rincon-Limas, D., Hoskins, R., Gnirke, A., Mungall, C., Wang, A., Kronmiller, B., Pacleb, J., Park, S., *et al.*, 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the drosophila genome. *Genome Biology*, **3**(12):research0086.1–0086.20.

Bernal, A., Crammer, K., Hatzigeorgiou, A., and Pereira, F., 2007. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS Computational Biology*, **3**(3):e54.

Blanchette, M., 2007. Computation and analysis of genomic multi-sequence alignments. *Annual Review of Genomics and Human Genetics*, **8**(1).

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**(4):708–715.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M., 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**(5611):1391–1394.

Bray, N. and Pachter, L., 2004. Mavid: Constrained ancestral alignment of multiple sequences. *Genome Res.*, **14**(4):693–699.

Brent, M., 2005. Genome annotation past, present, and future: How to define an orf at each locus. *Genome Research*, **15**(12):1777–1786.

Cox, T. and Cox, M., 2001. *Multidimensional Scaling*. Chapman & Hall/CRC.

Crollius, H. R., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Qutier, F., *et al.*, 2000. Estimate of human gene number provided by genome-wide analysis using tetraodon nigroviridis dna sequence. *Nat Genet*, **25**(2):235–238.

Crosby, M. A., Goodman, J. L., Strelets, V. B., Zhang, P., Gelbart, W. M., and Consortium, F., 2007. Flybase: genomes by the dozen. *Nucleic Acids Res*, **35**(Database issue):D486–D491.

Decaprio, D., Vinson, J. P., Pearson, M. D., Montgomery, P., Doherty, M., and Galagan, J. E., 2007. Conrad: Gene prediction using conditional random fields. *Genome Res*, .

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L., 1999. Improved microbial gene identification with glimmer. *Nucleic Acids Res*, **27**(23):4636–4641.

Eddy, S., 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol*, **3**(1):e10.

ENCODE Project Consortium, 2007. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, **447**(7146):799–816.

Frith, M., Bailey, T., Kasukawa, T., Mignone, F., Kummerfeld, S., Madera, M., Sunkara, S., Furuno, M., Bult, C., Quackenbush, J., *et al.*, 2006a. Discrimination of non-protein-coding transcripts from protein-coding mrna. *RNA Biol*, **3**(1).

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., and Grimmond, S. M., *et al.*, 2006b. The abundance of short proteins in the mammalian proteome. *PLoS Genet*, **2**(4):e52.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., and Couso, J. P., 2007. Peptides encoded by short orfs control development and define a new eukaryotic gene family. *PLoS Biol*, **5**(5):e106.

Gao, F. and Zhang, C.-T., 2004. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, **20**(5):673–681.

Gross, S. and Brent, M., 2006. Using multiple alignments to improve gene prediction. *J. Comput. Biol*, **13**:379–393.

Gross, S., Russakovsky, O., Do, C., and Batzoglou, S., 2007. Training conditional random fields for maximum parse accuracy. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 529–536. MIT Press, Cambridge, MA.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R., *et al.*, 2002. The genome sequence of the malaria mosquito anopheles gambiae. *Science*, **298**(5591):129–149.

Honeybee Genome Sequencing Consortium, 2006. Insights into social insects from the genome of the honeybee apis mellifera. *Nature*, **443**:931–949.

Jiggins, F. M. and Kim, K. W., 2007. A screen for immunity genes evolving under positive selection in drosophila. *J Evol Biol*, **20**(3):965–970.

Karlin, S., Chen, C., Gentles, A. J., and Cleary, M., 2002. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc Natl Acad Sci U S A*, **99**(26):17008–17013.

Kellis, M., Patterson, N., Birren, B., Berger, B., and Lander, E., 2004. Methods in comparative genomics: genome correspondence, gene identification and motif discovery. *J Comput Biol*, **11**:319–355.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E., 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**(6937):241–254.

Korf, I., Flicek, P., Duan, D., and Brent, M., 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics*, **17**(Suppl 1):S140–S148.

Lafferty, J., McCallum, A., and Pereira, F., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Liu, J., Gough, J., and Rost, B., 2006. Distinguishing protein-coding from non-coding rnas through support vector machines. *PLoS Genet*, **2**(4):e29.

Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Schwartz, A. S., Hou, M., *et al.*, 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1 *Genome Res*, **17**(6):760–774.

Margulies, E. H., Vinson, J. P., Program, N. I. S. C. C. S., Miller, W., Jaffe, D. B., Lindblad-Toh, K., Chang, J. L., Green, E. D., Lander, E. S., Mullikin, J. C., *et al.*, 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A*, **102**(13):4795–4800.

Meyer, I. and Durbin, R., 2002. Comparative ab initio prediction of gene structures using pair hmms. *Bioinformatics*, **18**(10):1309–1318.

Mignone, F., Grillo, G., Liuni, S., and Pesole, G., 2003. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res*, **31**(15):4639–4645.

Misra, S., Crosby, M., Mungall, C., Matthews, B., Campbell, K., Hradecky, P., Huang, Y., Kaminker, J., Millburn, G., Prochnik, S., *et al.*, 2002. Annotation of the drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol*, **3**(12):1–0083.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–562.

Nei, M. and Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol*, **3**(5):418–426.

Nekrutenko, A., Makova, K. D., and Li, W.-H., 2002. The ka/ks ratio test for assessing the protein-coding potential of genomic regions: An empirical and simulation study. *Genome Res.*, **12**(1):198–202.

Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W., and Guigo, R., 2003. Comparative gene prediction in human and mouse. *Genome Res.*, **13**(1):108–117.

Pedersen, J. and Hein, J., 2003. Gene finding with a hidden markov model of genome structure and evolution. *Bioinformatics*, **19**(2):219–227.

Reichhart, J.-M., 2005. Tip of another iceberg: Drosophila serpins. *Trends Cell Biol*, **15**(12):659–665.

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M. J., Chen, R., Meisel, R. P., *et al.*, 2005. Comparative genome sequencing of drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Res.*, **15**(1):1–18.

Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D. A., 2000. A drosophila complementary dna resource. *Science*, **287**(5461):2222–2224.

Saeys, Y., Rouze, P., and Van de Peer, Y., 2007. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. *Bioinformatics*, **23**(4):414.

Sarawagi, S. and Cohen, W., 2005. Semi-markov conditional random fields for information extraction. *Advances in Neural Information Processing Systems*, **17**:1185–1192.

Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., and Richards, S. e. a., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **15**(8):1034–1050.

Siepel, A. and Haussler, D., 2004. Computational identification of evolutionarily conserved exons. *Proceedings of the Eighth Annual International Conference on Resaerch in Computational Molecular Biology (RECOMB '05)*, :177–186.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., *et al.*, 2002a. A drosophila full-length cdna resource. *Genome Biol*, **3**(12):1–80.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., *et al.*, 2002b. The drosophila gene collection: Identification of putative full-length cdnas for 70% of d. melanogaster genes. *Genome Res.*, **12**(8):1294–1300.

Sutton, C. and McCallum, A., 2006. An introduction to conditional random fields for relational learning. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., *et al.*, 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**(6950):788–793.

Vinson, J. P., DeCaprio, D., Pearson, M. D., Luoma, S., and Galagan, J. E., 2007. Comparative gene prediction using conditional random fields. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1441–1448. MIT Press, Cambridge, MA.

Yang, Z., 1997. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**(5):555–556.

Yang, Z. and Bielawski, J., 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, **15**(12):496–503.

Yang, Z. and Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, **17**:32–43.

Yu, P., Ma, D., and Xu, M., 2005. Nested genes in the human genome. *Genomics*, **86**(4):414–422.

Zdobnov, E. M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R. R., Christophides, G. K., Thomasova, D., Holt, R. A., Subramanian, G. M., *et al.*, 2002. Comparative genome and proteome analysis of anopheles gambiae and drosophila melanogaster. *Science*, **298**(5591):149–159.

Zhang, L., Pavlovic, V., Cantor, C. R., and Kasif, S., 2003. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.*, **13**(6a):1190–1202.

Zhang, M. Q., 2002. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet*, **3**(9):698–709.

Figure Legends

Figure 1: Evolutionary distances relating 12 *Drosophila* species. (A) Phylogenetic tree and estimated neutral branch lengths for the species. Tree topology follows the accepted phylogeny of these species (*Drosophila* Comparative Genome Sequencing and Analysis Consortium, submitted). Neutral substitution rates estimated from 12,861 fourfold degenerate sites in syntenic one-to-one orthologs (see Supplemental Methods). (B) Pairwise distance of each of the 11 other *Drosophila* species from *D. melanogaster*, as compared to similarly estimated distances for vertebrates. (C) Total independent branch length provided by several subsets of the *Drosophila* species used to benchmark multi-species methods.

Figure 2: Overall discovery power of discriminative metrics using 12 genomes. (A) ROC curves showing sensitivity and specificity of each metric on classifying 10,722 known exons and 39,181 random non-coding regions. Comparative methods tended to outperform single-sequence metrics, with the exception of a baseline sequence conservation metric. CSF and the dN/dS test achieved near-perfect specificity, while RFC achieved high sensitivity. (B) Summary error statistics for each metric computed from the ROC curves. Minimum Average Error (MAE) is the minimum average of the false negative rate and false positive rate. Area Above the Curve (AAC) is the area above the ROC curve in the unit square. (C) MAE and AAC error statistics for each metric when the dataset is partitioned into several sequence length categories. All metrics tended to perform better on longer sequences than on shorter sequences. Comparative methods strongly outperformed single-sequence metrics on short sequences (60-240nt). Inset: relative size of each sequence length category.

Figure 3: Pairwise discovery power using different informant species. (A) ROC curves for K_A/K_S using *D. melanogaster* with each of five different informant species. Species at a wide range of evolutionary distances performed comparably, except for *D. erecta*, the most closely related to *D. melanogaster*, which clearly underperformed the others. (B) MAE and AAC error statistics for each pairwise comparative metrics applied to the same five informants. *D. ananassae* (blue) is overall the preferred informant, but not uniformly so. For TBLASTX, the performance is also shown using mosquito (*Anopheles gambiae*) and honeybee (*Apis mellifera*), which led to worse performance than the *Drosophila* species. No pairwise comparison outperformed the best single-sequence metric (Z curve).

Figure 4: Multi-species discovery power using increasing numbers of informant species. (A) ROC curves for the dN/dS test using subsets of *Drosophila* species corresponding to increasingly broad phylogenetic clades from *D. melanogaster* (see Figure 1). Discriminatory power steadily increased as more informants were used, leading to strictly better sensitivity and specificity. (B) MAE and AAC error statistics for each multi-species comparative metric using the same subsets of informants. Also shown for comparison are the best pairwise analysis and the bent single-sequence metric, both of which are outperformed by multi-species methods with sufficient informants. (C) Effect of additional species was most pronounced for short exon lengths. (x-axis) mean length within a quantile of the sequence length distribution (y-axis) sensitivity of the dN/dS test within each quantile at fixed specificity (99%).

Figure 5: Independence of metrics and discovery power of metric combinations. (A) ROC curves showing the performance of two hybrid metrics created by combining five comparative and single-sequence metrics using Linear Discriminant Analysis (LDA) or a Support Vector Machine (SVM). The hybrid metrics outperformed all of their input metrics. (B) Multidimensional scaling (MDS) visualization in which each point represents a metric and the distance between any two points approximately represents their dissimilarity, measured as 1-(rank correlation of the scores of the known exons). Hybrid metrics appear closer to the center, suggesting that they successfully combine distinct information from the individual metrics.

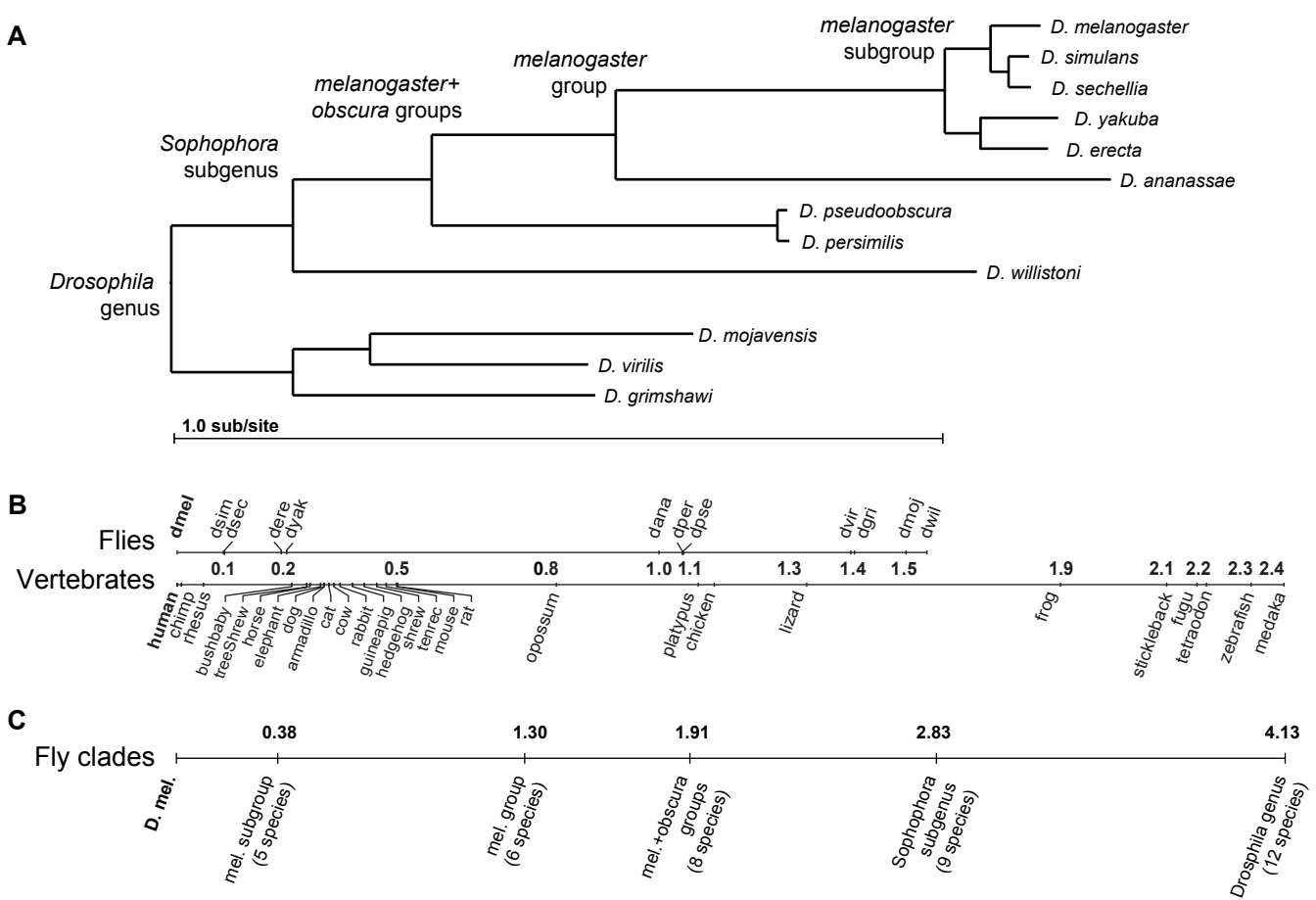

A
melanogaster subgroup
D. melanogaster
D. simulans
D. sechellia
D. yakuba
D. erecta
melanogaster group
melanogaster+ obscura groups
Sophophora subgenus
D. ananassae
D. pseudoobscura
D. persimilis
D. willistoni
Drosophila genus
D. mojavensis
D. virilis
D. grimshawi
1.0 sub/site
B
Flies
dmel
dsim
dsec
dere
dyak
dana
dper
dpse
dvir
dgri
dmoj
dwil
Vertebrates
0.1
0.2
0.5
0.8
1.0
1.1
1.3
1.4
1.5
1.9
2.1
2.2
2.3
2.4
human
chimp
rhesus
bushbaby
treeShrew
horse
elephant
dog
armadillo
cat
cow
rabbit
guineapig
hedgehog
shrew
tenrec
mouse
rat
opossum
platypus
chicken
lizard
frog
stickleback
fugu
tetraodon
zebrafish
medaka
C
Fly clades
0.38
1.30
1.91
2.83
4.13
D. mel.
mel. subgroup (5 species)
mel. group (6 species)
mel.+obscura groups (8 species)
Sophophora subgenus (9 species)
Drosphila genus (12 species)

Figure 1

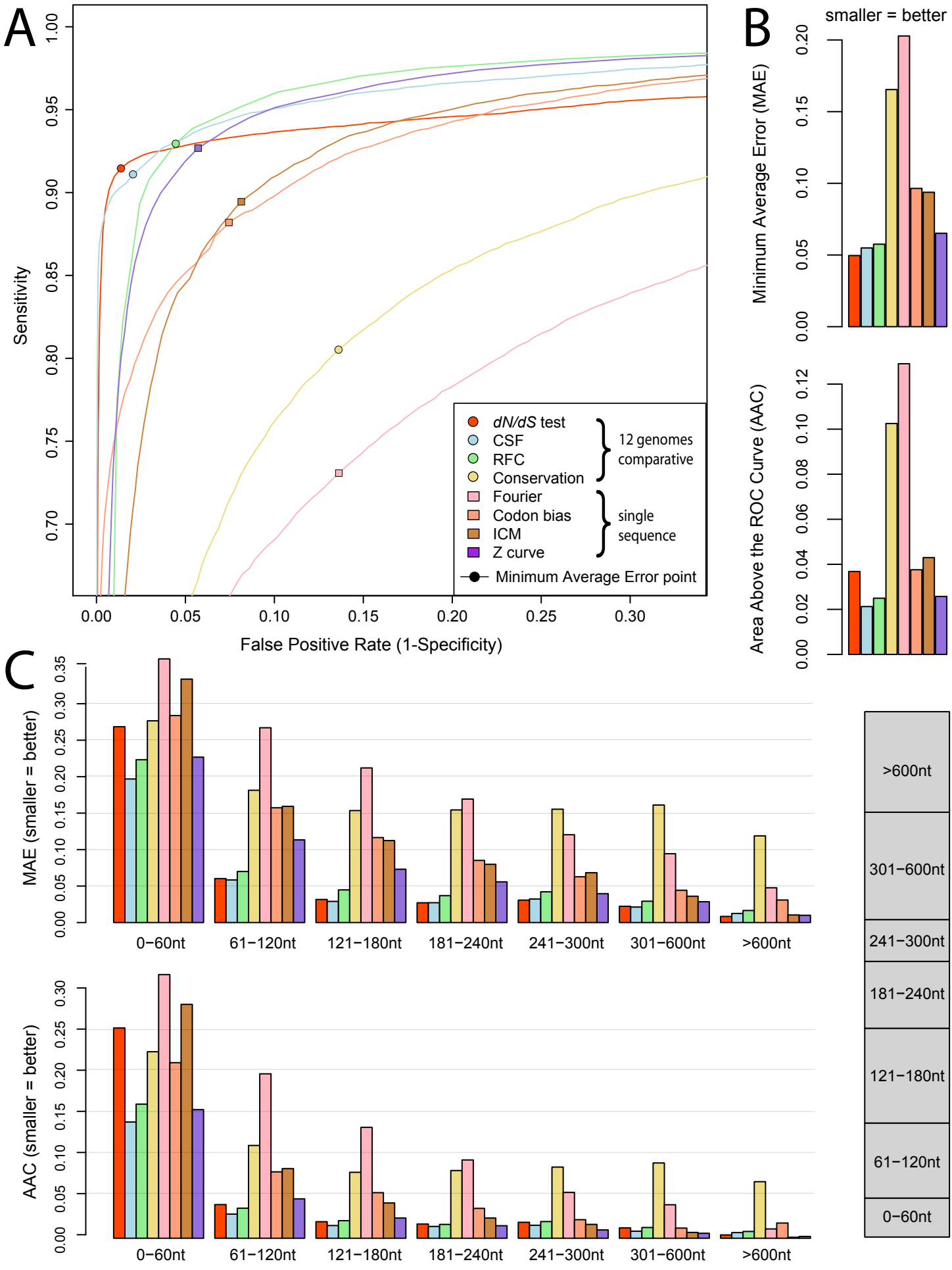

A
Sensitivity
False Positive Rate (1-Specificity)
dN/dS test
CSF
RFC
Conservation
Fourier
Codon bias
ICM
Z curve
12 genomes comparative
single sequence
Minimum Average Error point
B
smaller = better
Minimum Average Error (MAE)
Area Above the ROC Curve (AAC)
C
MAE (smaller = better)
AAC (smaller = better)
0−60nt
61−120nt
121−180nt
181−240nt
241−300nt
301−600nt
>600nt

Figure 2

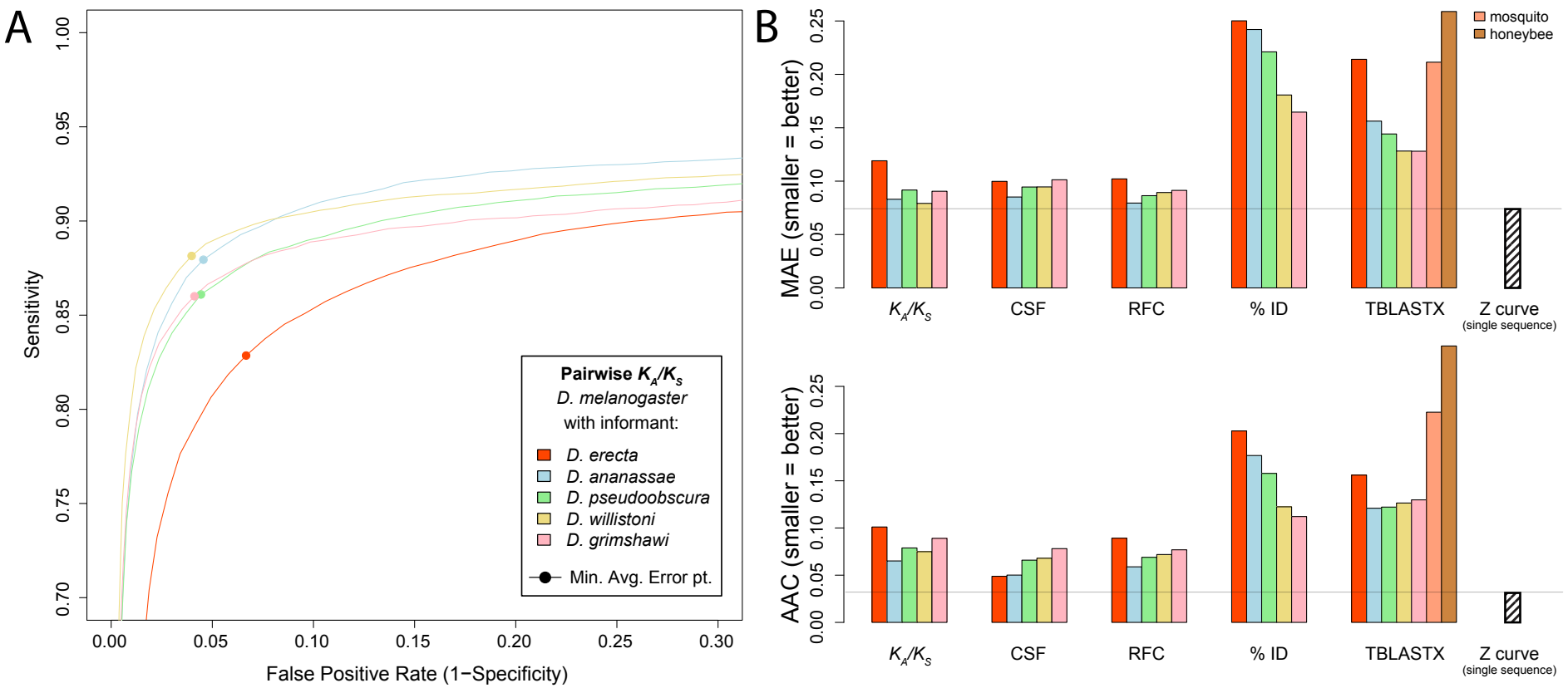

A
Sensitivity
False Positive Rate (1−Specificity)
0.70
0.75
0.80
0.85
0.90
0.95
1.00
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Pairwise K_A/K_S
D. melanogaster
with informant:
D. erecta
D. ananassae
D. pseudoobscura
D. willistoni
D. grimshawi
Min. Avg. Error pt.
B
MAE (smaller = better)
AAC (smaller = better)
mosquito
honeybee
K_A/K_S
CSF
RFC
% ID
TBLASTX
Z curve
(single sequence)

Figure 3

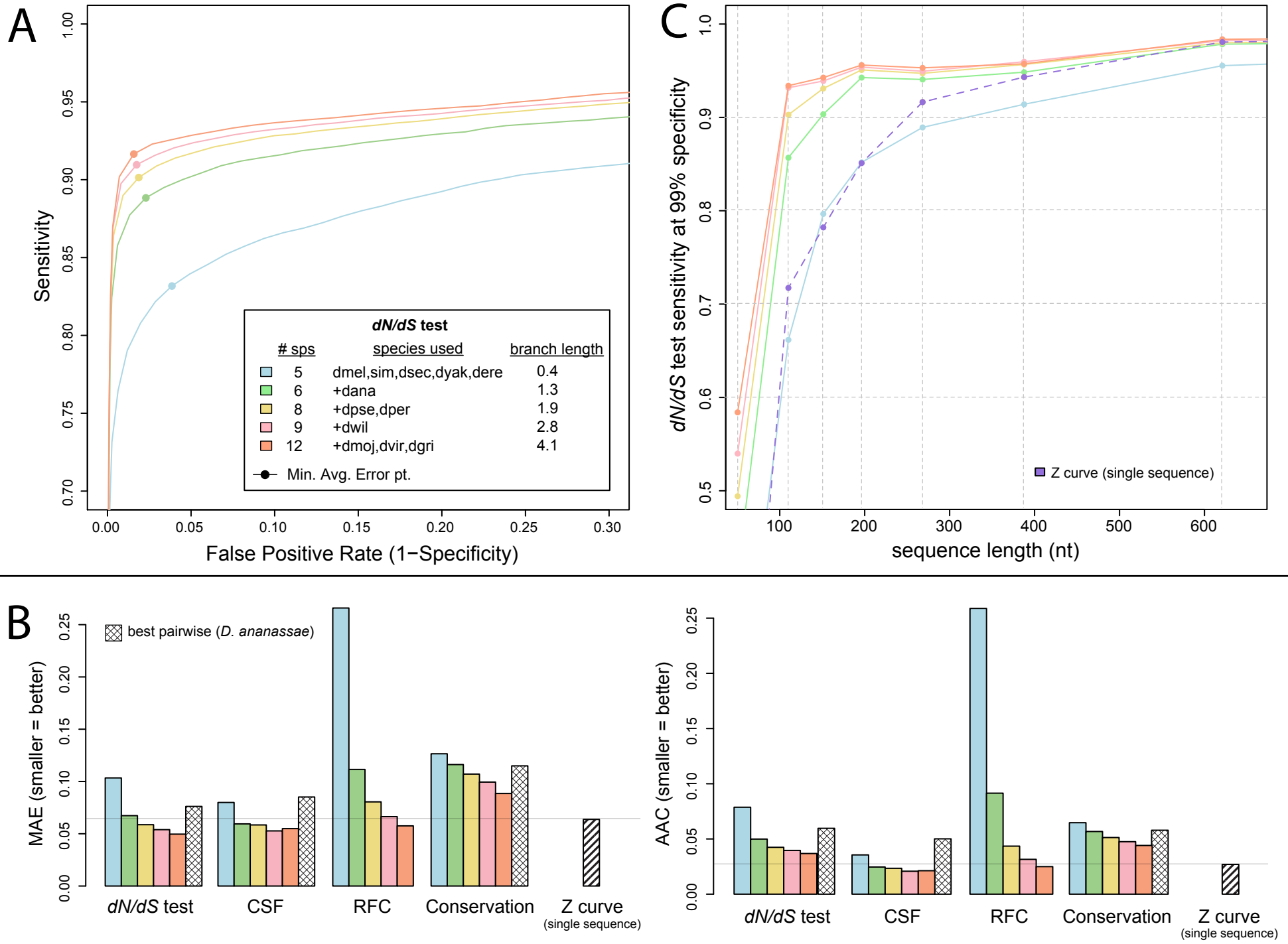
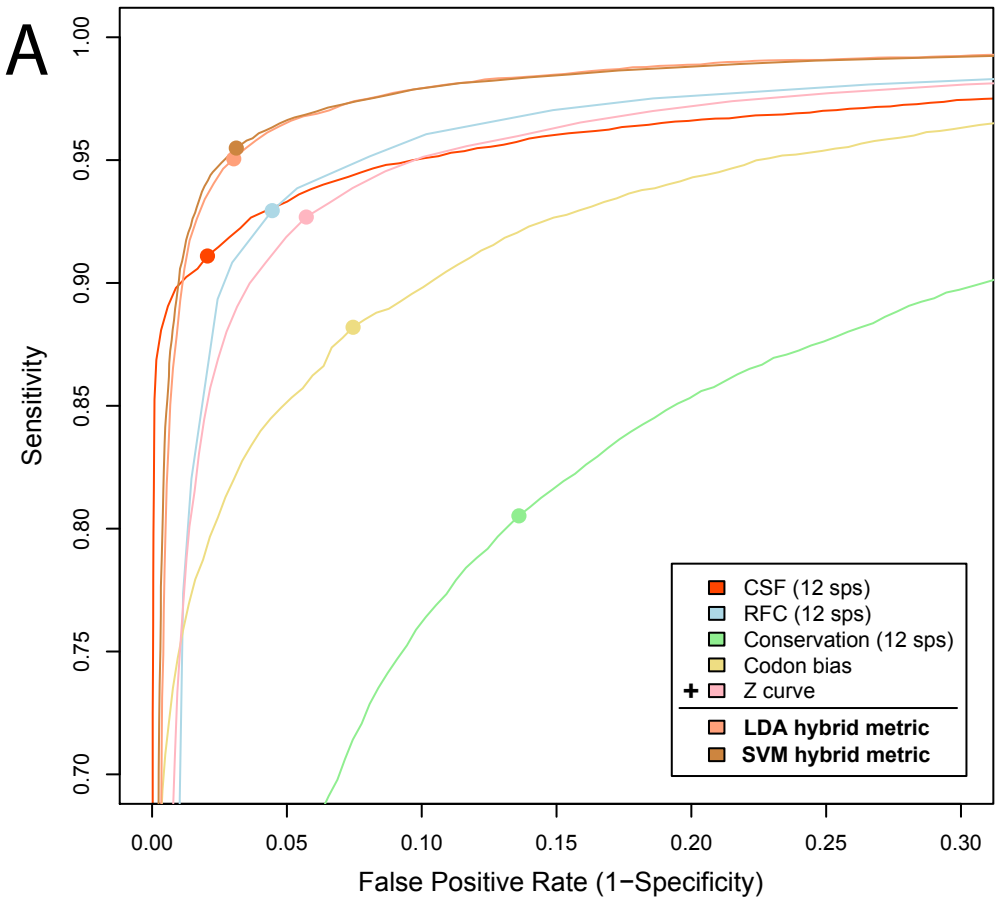

A
Sensitivity
False Positive Rate (1−Specificity)
dN/dS test
sps
species used
branch length
5
dmel,sim,dsec,dyak,dere
0.4
6
+dana
1.3
8
+dpse,dper
1.9
9
+dwil
2.8
12
+dmoj,dvir,dgri
4.1
Min. Avg. Error pt.
C
dN/dS test sensitivity at 99% specificity
sequence length (nt)
Z curve (single sequence)
B
MAE (smaller = better)
best pairwise (D. ananassae)
dN/dS test
CSF
RFC
Conservation
Z curve
(single sequence)
AAC (smaller = better)
dN/dS test
CSF
RFC
Conservation
Z curve
(single sequence)

Figure 4


A
Sensitivity
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.00
0.05
0.10
0.15
0.20
0.25
0.30
False Positive Rate (1−Specificity)
CSF (12 sps)
RFC (12 sps)
Conservation (12 sps)
Codon bias
Z curve
LDA hybrid metric
SVM hybrid metric

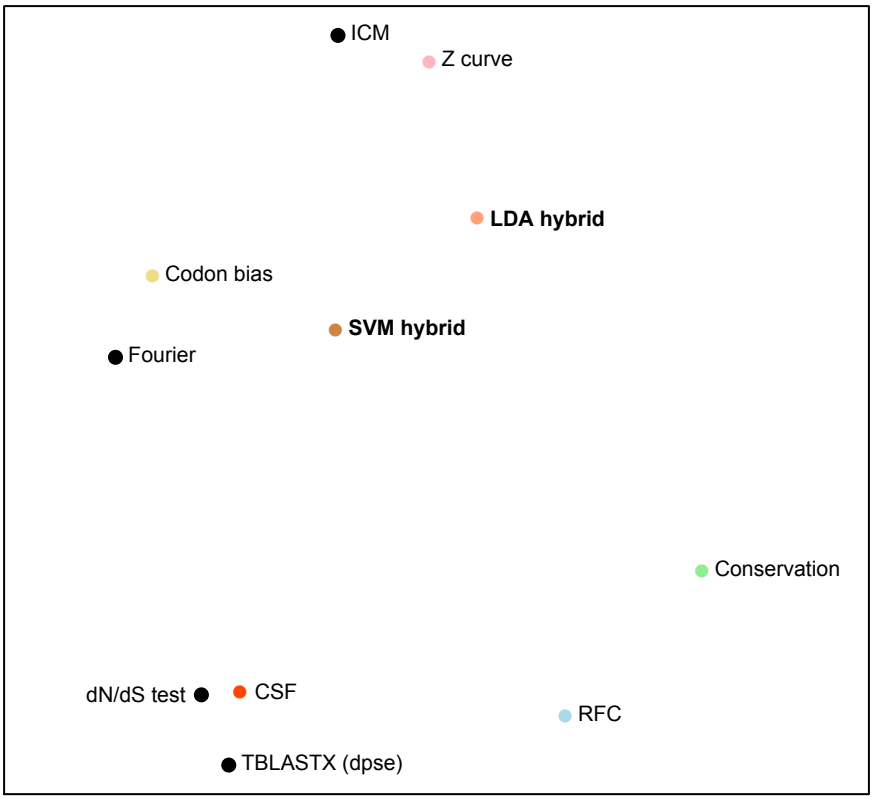

B
ICM
Z curve
LDA hybrid
Codon bias
SVM hybrid
Fourier
Conservation
dN/dS test
CSF
RFC
TBLASTX (dpse)

Figure 5