LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-409564

nGASP - the nematode genome annotation assessment project

A. Coghlan, T. J. Fiedler, S. J. McKay, P. Flicek, T. W. Harris, D. Blasiar, J. Allen, L. D. Stein

December 23, 2008

nGASP - the nematode genome annotation assessment project

Disclaimer

nGASP – the nematode genome annotation assessment project

Avril Coghlan[1*], Tristan J. Fiedler[2*], Sheldon J. McKay[3], Paul Flicek[4], Todd W. Harris[3], Darin Blasiar[5], the nGASP Consortium, and Lincoln D. Stein[3,§]

[*]These authors should be considered as joint first authors.

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

[2]Department of Biological Sciences, Florida Institute of Technology, Melbourne, FL 32901

[3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

[4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

[5]Washington University School of Medicine, St Louis, MO 63108

[§]Corresponding author

Email addresses:

TJF: fiedler@fit.edu

DB: dblasiar@watson.wustl.edu

AC: alc@sanger.ac.uk

PF: flicek@ebi.ac.uk

TWH: toddwharris@gmail.com

SJM: mckays@cshl.edu

LDS: lstein@cshl.edu

Note:

In 2008 the article-processing charge is £950 (\u20ac1195, US$1855) for *BMC Bioinformatics*, free if submitting author's institute is a member of BMC.

FIGURES: Figures must be submitted as separate image files. EPS preferred for diagrams, PNG for photos/images. Fig. Legends should be after refs. For each figure, the following information should be provided: Figure number (in sequence, using Arabic numerals - i.e. Figure 1, 2, 3 etc); short title of figure (maximum 15 words); detailed legend, up to 300 words. Individual figure files should not exceed 10 MB.

Figures on the web:

* width of 600 pixels (standard), 1200 pixels (high resolution).

Figures in the final PDF version:

* width of 85 mm for single column;

* width of 176 mm for double column;

* maximum height of 230 mm for figure and legend;

* image resolution should be approximately 300 dpi (dots per inch) at the final size.

Illustrations should be designed such that all information is legible at these dimensions. All lines should be wider than 0.5 pt when constrained to standard figure widths.

BioMed Central recommends the use of either Arial or Helvetica fonts using 12pt characters for text within figures. Vector figures should if possible be submitted as PDF files, which are usually more compact than EPS files.

TABLES: Each table should be numbered in sequence using Arabic numerals (i.e. Table 1, 2, 3 etc.). Tables should also have a title that summarizes the whole table, maximum 15 words. Detailed legends may then follow, but should be concise. Columns and rows of data should be made visibly distinct by ensuring the borders of each cell display as black lines. Commas should not be used to indicate numerical values. Color and shading should not be used.

Abstract

Background

While the *C. elegans* genome is extensively annotated, relatively little information is available for other *Caenorhabditis* species. The nematode genome annotation assessment project (nGASP) was launched to objectively assess the accuracy of protein-coding gene prediction software in *C. elegans*, and to apply this knowledge to the annotation of the genomes of four additional *Caenorhabditis* species and other nematodes. Seventeen groups worldwide participated in nGASP, and submitted 47 prediction sets for 10 Mb of the *C. elegans* genome. Predictions were compared to reference gene sets consisting of confirmed or manually curated gene models from WormBase.

Results

The most accurate gene-finders were 'combiner' algorithms, which made use of transcript- and protein-alignments and multi-genome alignments, as well as gene predictions from other gene-finders. Gene-finders that used alignments of ESTs, mRNAs and proteins came in second place. There was a tie for third place between gene-finders that used multi-genome alignments and *ab initio* gene-finders. The median gene level sensitivity of combiners was 78% and their specificity was 42%, which is nearly the same accuracy as reported for combiners in the human genome. *C. elegans* genes with exons of unusual hexamer content, as well as those with many exons, short exons, long introns, a weak translation start signal, weak splice sites, or poorly conserved orthologs were the most challenging for gene-finders.

Conclusions

This experiment establishes a baseline of gene prediction accuracy in *Caenorhabditis* genomes, and has guided the choice of gene-finders for the annotation of newly sequenced genomes of *Caenorhabditis* and other nematode species. We have created new gene sets for *C. briggsae*, *C. remanei*, *C. brenneri*, *C. japonica*, and *Brugia malayi* using some of the best-performing gene-finders.

Background

The promise of comparative genomics among the nematodes has motivated sequencing in *Caenorhabditis elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, and *C. japonica* [1-3]. While the *C. elegans* genome has been extensively annotated, relatively little information is available for the other *Caenorhabditis* genomes [4]. In addition, the genome of the distantly related nematode *Brugia malayi* was recently published [5], and those of many other nematodes are currently being sequenced such as *Pristionchus*, *Haemonchus*, *Meloidogyne*, and *Trichinella*. An essential step in the analysis of these genomes will be to identify and annotate their protein-coding genes, but it is not known which gene prediction systems perform best on nematode genomes. To address this issue, the nematode genome annotation assessment project (nGASP) was launched to assess the accuracy of protein-coding gene prediction software in *C. elegan*s, and then to apply this knowledge to annotating other *Caenorhabditis* genomes.

The nGASP project parallels recent computational prediction initiatives including CASP for protein structure prediction [6], GASP for *Drosophila* gene prediction [7], and EGASP for human gene prediction [8]. Scientists working in the

field of computational gene prediction were invited to participate in nGASP. Participants were provided with training and test sets, each comprising ten non-overlapping 1-Mb genomic sequence regions, representing ~10% of the *C. elegans* genome. We also provided auxiliary data to the participants to use for training their gene-finders, and producing the final predictions on the test regions. The auxiliary data included multi-genome alignments between *C. elegans*, *C. briggsae* and *C. remanei*, and alignments of ESTs, mRNAs and proteins to the *C. elegans* genome.

nGASP was conducted in two phases. The first phase of the competition was open to all gene prediction programs and was divided into three categories: category 1 predictions were based on genomic sequence alone (*ab initio* gene-finders); category 2 used nucleotide level multi-genome alignments; and category 3 predictions used alignments of expressed sequences such as proteins, ESTs, and assembled mRNAs. After the first phase of the competition was complete, we posted the output of each of the predictors to the nGASP web site (http://dev.wormbase.org/ngasp). We then began phase two of the competition, which was open to 'combiners' (category 4), defined as gene prediction systems that use gene models created by other annotation software, and any of the data used as input for the phase one gene-finders. To assess the accuracy of the submitted predicted gene sets, we quantified their sensitivity and specificity in predicting coding regions by using the metrics from GASP [7] and EGASP [8]. Here, we describe the performance of the most accurate gene-finders in *C. elegans*, identify some common features of *C. elegans* genes that the majority of gene-finders find hard to predict correctly, and discuss the choice of gene predictors for the annotation of the newly sequenced genomes of other nematode species.

Results and Discussion

Submitted Gene Sets

Seventeen groups worldwide participated in nGASP, and submitted 47 prediction sets for 10 Mb of the *C. elegans* genome (Table 1). Several groups submitted predicted gene sets for more than one category, or more than one entry per category generated by running their programs under different parameter sets. The submitted gene sets, and the details of the parameters used to make them, are available on the nGASP ftp site ftp://ftp.wormbase.org/pub/wormbase/nGASP/.

Procedure for Evaluation of Gene-Finding Accuracy

The 10-Mb of test DNA sequence consisted of ten non-overlapping 1-Mb genomic regions of the *C. elegans* genome (Table 2). The gene predictions submitted to nGASP were evaluated using two reference gene sets drawn from WormBase (release WS160), an intensively curated gene prediction set: (i) **ref1,** a 'sensitivity/accuracy' set consisting of genes from the test regions that were supported by full-length cDNAs, and (ii) **ref2**, a 'full set' that contained all curated genes from the test regions (see **Methods**). We assessed sensitivity (Sn) using the **ref1** reference and specificity (Sp) using the **ref2** reference. nGASP differed from the *Drosophila* GASP [7] and human EGASP [8], in that curated gene structures for *C. elegans* were already publicly available, but participants were requested to not consult WormBase, GenBank or other databases for the curated gene models in the test regions.

For each submitted gene set, we assessed its ability to accurately predict protein coding regions at the base, exon, isoform and gene levels, following the definitions of EGASP [8]. The least stringent metrics were base level sensitivity and specificity, which measure whether a gene predictor is able to correctly classify a base

as coding. By ‘exon’, we mean the protein-coding part of an exon (also known as the CDS, or coding sequence). Exon level metrics measure the ability of a gene prediction system to identify the exact left and right borders of the protein-coding regions of exons in the reference sets. Isoform level accuracy is the most stringent test. One *C. elegans* gene can produce several alternative spliced transcripts. For the purposes of nGASP we considered only the protein-coding portion of a transcriptional isoform, and scored a correctly predicted isoform if the protein-coding portions of all its exons were predicted accurately and no extra full or partially protein-coding exons were predicted. The gene level assessment of accuracy was intermediate in stringency between the exon and isoform levels. To be scored correct at the gene level, a gene predictor had to call at least one of the gene's isoforms correctly.

Results from Evaluation of Gene-Finding Accuracy

The best submitted gene prediction sets had base level sensitivity in excess of 99% and specificity of more than 93% (Table 3; Figure 1). This means that the best gene predictors are able to identify almost all the protein-coding bases in the *C. elegans* genome and only occasionally predict that a non-coding base is coding. At the exon level, the best submitted gene sets had sensitivities of more than 91% and specificities of more then 83%. Although most gene-finders identify most true coding bases correctly, they often do misidentify the boundaries of protein-coding exons. At the base and exon levels, specificities were lower than sensitivities. This may reflect a number of inaccurate gene models in the **ref2** gene set, which included gene models not fully supported by transcript evidence, and perhaps also reflects exons that are missing from the **ref2** gene set.

The ultimate goal of a gene predictor is to predict entire genes correctly, including every alternative isoform. However, in practice gene-finders do not predict alternative isoforms of a gene very well. At the isoform level, the best gene sets had sensitivities of about 66% and specificities of about 56%. That is, the best gene-finders each missed about 34% of true *C. elegans* isoforms, indicating that gene-finders still need improvements in predicting alternative splice forms. At the gene level, the best submitted gene sets had sensitivities and specificities in excess of 80% and 58% respectively. That is, for 80% of genes in the **ref1** reference set, the best gene predictors called at least one splicing isoform correctly across the entire length of its protein-coding region.

The isoform level is the most stringent level of assessment. However, given the low success of most gene-finders for predicting alternative splicing, gene level accuracy is generally considered more important for the purpose of annotating a newly sequenced genome such as that of *Caenorhabditis remanei*. That is, it is considered more important to predict at least one isoform of each gene correctly, rather than to predict all isoforms of one gene correctly and no isoforms of a second gene correctly. At the gene level, the most accurate gene-finders were combiners (Figure 1). Gene predictors that use alignments of ESTs, mRNAs and proteins came in second place. Combiners had higher sensitivity than algorithms that used expressed sequence alignments at the gene level (medians: combiners 78%, expressed sequence-based 68%, P=0.04). However, in terms of specificity, there was no significant difference in gene level accuracy between combiners and gene predictors that used transcript and protein alignments (medians: combiners 42%, expressed sequence-based 39%, P=0.1). Thus, by using diverse data such as expressed sequence alignments, multi-genome alignments and gene sets from different gene-finders,

combiners improved the sensitivity of their predictions above those based on expressed sequence alignments alone. This agrees with EGASP [8], which reported that combiners had higher gene level sensitivities for human genes compared to gene-finders that used expressed sequence alignments alone (medians: combiners 70%, expressed sequence-based 64%) [8].

At the gene level, prediction algorithms that used expressed sequence alignments had higher sensitivity than *ab initio* gene predictors (medians: expressed sequence-based 68%, *ab initio* 54%, P=0.05), as well as higher specificity (expressed sequence-based 39%, *ab initio* 32%, P=0.01). This demonstrates that use of expressed sequence data leads to considerable improvements in the accuracy of gene-finders for *C. elegans*. This mirrors the findings of EGASP, which also reported higher gene-sensitivities for gene-finders that used transcript or protein alignments compared to *ab initio* gene-finders (medians: expressed sequence-based 63%, *ab initio* 18%), as well as higher gene-specificities (expressed sequence-based 55%, *ab initio* 8%) [8].

There was a tie for third place between gene prediction algorithms that used multi-genome alignments and *ab initio* gene-finders. The addition of multi-genome alignments to *C. briggsae* and *C. remanei* gave no statistically significant improvement in accuracy over *ab initio* predictions. This was surprising, as the EGASP project reported that gene-finders that used multi-genome alignments were more accurate than *ab initio* gene-finders for predicting human genes, in terms of both gene level sensitivities (medians: *ab initio* 18%, multi-genome 26%) and specificities (*ab initio* 8%, multi-genome 19%). This may reflect the relatively high gene level accuracy of *ab initio* gene-finders in *C. elegans* (medians: gene Sn 54%, Sp 32%), compared to human (Sn 18%, Sp 8%) [8], probably due to the compact nature of the *C. elegans* genome. In addition, it is possible that the evolutionary distances

separating *C. elegans*, *C. briggsae* and *C. remanei* are less suited for inference of protein coding genes from multi-genome alignments than the corresponding set of vertebrate genomes used in the EGASP study. Furthermore, a difference in the way that the reference sets were defined for nGASP and EGASP could contribute to the observed difference in accuracy. For example, because nGASP used different reference sets to estimate sensitivity and specificity, this might lead to different results compared to EGASP, which relied on a single set of reference genes to calculate both sensitivity and specificity.

In both nGASP and EGASP, the best gene-finders were combiners. However, in nGASP the median gene level sensitivity of combiners was 78% and specificity was 42%, while in EGASP the median gene level sensitivity of combiners was 70% and specificity was 52% [8]. In *C. elegans*, about 8% more of the true genes are predicted correctly, but 10% fewer of the gene predictions made are structurally correct. The lower specificity in *C. elegans* suggests that there are more real isoforms and/or real genes missing from the *C. elegans* curated gene set, compared to the human curated gene set. This could be due to the far smaller amount of transcript data available for *C. elegans* and/or more conservative manual curation of weakly supported isoforms or genes by WormBase. Using the average of the sensitivity and specificity as an overall metric of accuracy, the *C. elegans* combiner gene sets were slightly less accurate (median 59%) than the human combiner gene sets (median 61%).

Factors Affecting Gene-Finding Accuracy

To understand which factors affect the accuracy of gene-finders in *C. elegans*, we identified features of genes that were not predicted correctly by the *ab initio* gene-

finders, gene-finders that used multi-genome alignments, and gene-finders that used expressed sequence alignments. The percentage of gene sets in which a true gene was predicted correctly (using the **ref1** reference gene set) was found to be correlated with nine features of genes (Figure 2):

(i) the lowest 'hexamer score' of any of the exons in the gene (Spearman's ρ=0.38, $P<10^{-16}$), using the score based on the frequency of 6-bp words from Genefeatures in the AceDB software [9],

(ii) the number of exons in the gene (ρ=-0.36, $P<10^{-16}$),

(iii) the length of the shortest exon in the gene (ρ=0.30, $P=10^{-11}$),

(iv) the length of the longest intron in the gene (ρ=-0.29, $P=10^{-9}$),

(v) the strength of the translation start signal (ρ=0.28, $P=10^{-9}$), as measured by Genefeatures,

(vi) the lowest score of any of splice sites in the gene (ρ=0.25, $P=10^{-7}$), as measured by Genefeatures,

(vii) the percent identity with the *C. briggsae* ortholog at the amino acid level (ρ=0.22, $P=10^{-5}$), based on an alignment from the TreeFam database of gene families [10],

(viii) the maximum distance to a neighbouring gene (ρ=-0.16, P=0.0003), and

(ix) the number of isoforms in the gene (ρ=-0.11, P=0.02).

That is, the *C. elegans* genes that are hardest for gene-finders to predict correctly are those with an exon of unusual hexamer content, lots of exons, a very short exon, a very long intron, a weak translation start signal, a weak splice site, a poorly conserved ortholog, far from one of its neighbouring genes, or with many isoforms. We suggest that developers of gene-finding programs should concentrate on improving accuracy on these types of genes. The correlation with these features tended to be stronger for

ab initio gene-finders than expressed sequence-based gene-finders (Figure 2). For example, the correlation with the lowest hexamer score for the exons in a gene was higher for *ab initio* gene-finders than for expressed sequence-based gene-finders (ρ=0.42 and 0.22, Z-test: P=0.0005). We observed weak or nonexistent correlations with other features that we examined, such as the length of the longest exon in a gene (P>0.05), length of the shortest intron (P>0.05), whether the adjacent genes are on the same strand (P>0.05), existence of embedded genes with a gene's introns (ρ=-0.11, P=0.01), whether a gene is member of an operon (P>0.05), whether neighbouring genes are paralogs (inferred from TreeFam [10]; P>0.05), and whether the gene overlaps a simple repeat or transposable element (P>0.05).

There were 19 genes that were missed in all of the category 1, 2 and 3 gene sets, which must be the most difficult-to-predict: *C06G3.7* (*trxr-1*), *C08G5.5*, *C33H5.14* (*ntp-1*), *C55F2.1*, *D1009.1*, *F18E9.3*, *F57F5.2* (*gcy-33*), *R04E5.7*, *R04E5.8*, *T07D3.4*, *T07F12.4*, *Y105E8A.7* (*eat-18*), *Y43H11AL.1*, *Y43G8AL.7, Y54E5B.1* (*smp-1*), *Y55F3BR.6*, *ZC455.6* , *ZC477.1* (*ssq-3*), and *ZC8.4* (*lfi-1*). Several of these genes have unusually long introns of >1400 bp (*Y43H11AL.1*, *Y43G8AL.7*, *Y54E5B.1*), unusually short exons of <40 bp (*D1009.1*, *Y105E8A.7*, *ZC8.4*), poorly conserved orthologs (*C08G5.5*, *R04E5.7*, *R04E5.8*), lots of exons (*F18E9.3, T07D3.4*), or are very far from one of their neighbours (*T07F12.4*).

New Gene Sets for *C. remanei*, *C. brenneri, C. japonica* and *Brugia malayi*

To judge which gene-finders in each category performed best, we used the average of the gene level sensitivity and specificity for a gene set as a metric of overall accuracy. In collaboration with several of the nGASP contributors, we are assembling new gene sets for *C. elegans*, *C. briggsae*, *C. brenneri*, *C. remanei*, *C. japonica* and *Brugia*

malayi using the three best performing of the gene-finders that used transcript/protein alignments: MGENE (Schweikert et al, submitted), AUGUSTUS [11] and FGENESH [12]. The best performing combiner, JIGSAW [13], is being used to combine the MGENE, AUGUSTUS and FGENESH predictions into a single nGASP gene set for each species that will form the basis of curated gene sets for the new genomes and will be used to improve curated gene models in *C. elegans*. All gene sets will be available from ftp://ftp.wormbase.org/pub/wormbase/nGASP_gene_predictions/predictions/ and will also be displayed in the genome browsers for these species at http://www.wormbase.org.

Conclusions

This experiment establishes a baseline of gene prediction accuracy in *Caenorhabditis* genomes, and is guiding the choice of gene prediction systems for the annotation of newly sequenced genomes for *Caenorhabditis* and other nematode species. At present, combiners are more accurate than other classes of gene prediction algorithm in *C. elegans*. However, the accuracy of the combiners would presumably benefit by increasing the accuracy of the component gene prediction sets that they are given. We have also identified features of *C. elegans* genes that are difficult to predict for *ab initio* gene-finders and gene-finders that use transcript-, protein- and multi-genome alignments, and hope that leaders in the gene prediction field will rise to the challenge of improving accuracy on such genes.

Methods

Data Provided to the nGASP Participants

Genomic DNA Sequence: To select the nGASP test and training regions, we divided the WormBase WS160 *C. elegans* genome sequence into 102 non-overlapping regions of 1 Mb, and discarded regions of less than 1 Mb from the high-coordinate ends of the six chromosomes, leaving a set of 96 1 Mb regions. Representative training and test regions were selected from these regions based on gene density and conservation, following the strategy used to select the human ENCODE regions [8]. We measured gene density in each region by counting the number of curated genes, and assessed conservation with *C. briggsae* by using the number of bases covered by strong WABA [14] matches to *C. briggsae*. Regions were classified as having high or low gene density or conservation if their values lay in the top or bottom 33% percentiles respectively. The test and training sets each consisted of ten 1 Mb regions that were randomly chosen from the sets of regions with particular combinations of high/low gene density and high/low conservation (for example, we randomly chose two of the high conservation, low gene density autosomal regions; Table 2).

***Auxiliary Training Data*:** We requested that gene-finders that had previously been trained using a large fraction of *C. elegans* confirmed genes or other data outside the supplied training sets be retrained solely on the training set provided by the nGASP project, namely:

(i) the coordinates of repeats found by RepeatMasker (A. Smit, unpublished, http://www.repeatmasker.org) in the training regions.

(ii) the coordinates of coding exons, introns and UTRs in 584 confirmed isoforms (CDSs) of 432 genes in the training regions. An isoform was considered as confirmed if it was supported from start to end by mRNA, EST or OST transcript data.

(iii) the coordinates of coding exons, introns, and UTRs in 1583 'unconfirmed' isoforms of 1461 genes in the training regions. These genes lacked any confirmed isoforms.

(iv) the DNA sequence for the 'cb1' assembly of the *C. briggsae* genome.

(v) the DNA sequence for the 'pcap2' assembly of the *C. remanei* genome.

(vi) a multi-genome alignment between *C. elegans*, *C. briggsae* and *C. remanei* for the *C. elegans* training regions, made using MLAGAN version 1.21 [15].

(vii) the amino acid sequences of 42,496 proteins that have BLAST [16] matches to the test or training regions, excluding matches to proteins encoded by genes in the test regions. The BLAST matches were made by running BLAST with an *E*-value cut-off of 0.1 against proteins from *C. elegans* (wormpep160), *C. briggsae* (brigpep160), *Drosophila melanogaster* (FlyBase [17]), *Saccharomyces cerevisiae* (SGD [18]), UniProt [19], and human (Ensembl [20] and RefSeq [21]).

(viii) the nucleotide sequences of 20,141 *C. elegans* ESTs/cDNAs that have BLAT matches to the test or training regions.

(ix) the coordinates of the BLAST and BLAT matches in (vii) and (viii) in the test and training regions.

Participants were allowed to use different data for training, and for making predictions in the test regions, according to the nGASP category under which they were submitting a gene prediction set. The repeat sequences (i) and genes in the training regions (ii and iii) could be used by all participants. Participants who submitted *ab initio* (category 1) gene sets were not allowed to use any additional data

for training or making gene sets. For gene-finders that used multi-genome alignments (category 2), participants could use the *C. briggsae* and *C. remanei* assemblies (iv and v) and the MLAGAN multi-genome alignment (vi). They also were allowed to generate a different multi-genome alignment using the tool(s) of their choice. For gene predictors that used expressed sequence alignments (category 3), participants could use the protein and transcript matches (vii, viii, ix), or they could choose a different alignment algorithm to realign the protein and transcript sequences contained in these sets.

For combiners (category 4), participants could use any of the auxiliary data allowed for categories 1-3, as well as the gene predictions submitted for categories 1 through 3 during nGASP phase one. Category 4 participants were also supplied with the coordinates of coding exons, introns, and UTRs in 386 confirmed isoforms of 242 genes in the 5' halves of each of the phase one test regions, which could be used as an additional training set. Because of this, combiners were evaluated using **ref1** and **ref2** gene sets drawn from the 3' halves of each phase one test region.

Submission of Gene Sets

The submitted gene prediction files were required to be in GFF3 format (L. Stein, unpublished; http://song.sourceforge.net/gff3.shtml), an extension of GFF (Gene Feature Format; R. Durbin and D. Haussler; http://www.sanger.ac.uk/Software/GFF). The GFF3 files were required to contain lines for gene, mRNA, CDS, and 5' and 3' UTR features. The format of gene prediction files submitted to nGASP was validated using a GFF3 format validator (P. Canaran, unpublished; http://dev.wormbase.org/db/validate_gff3/validate_gff3_online).

Resources for Assessing Predictions: the Reference Gene Sets

Predictions were compared to two different reference gene sets based on data in WormBase WS160 [4]: (i) all confirmed isoforms in the test regions ('**ref1**'), and (ii) all isoforms in all genes in the test regions ('**ref2**'). **Ref1** consisted of 605 isoforms from 493 different genes, and **ref2** consisted of 2250 isoforms from 1956 different genes. For phase two, we evaluated combiners using the 3' halves of each test region. The phase two **ref1** and **ref2** reference sets contained 313 isoforms from 249 different genes, and 1130 isoforms from 966 different genes, respectively.

We used **ref1** to assess sensitivity and **ref2** to assess specificity. This is because the true-positive and false-negative counts calculated by comparison to **ref1** are more reliable than those calculated using **ref2**, as the gene models in **ref1** are of higher quality. In contrast, the false-positive counts calculated by comparison to **ref2** are more reliable, because a higher fraction of true genes are represented by gene models in **ref2**.

Evaluation of Accuracy of Submitted Gene Sets

Two sets of evaluation software were written for nGASP. The first (P. Flicek, unpublished) was based on the earlier EGASP [8] evaluation software, but was extended to handle the GFF3 format for nGASP. The second software (A. Coghlan, unpublished) was written independently but calculated the same accuracy statistics.

Data Availability and Visualisation

The nGASP test and training data, the submitted gene predictions and the command-line options and parameters used to generate them, and the **ref1** and **ref2** reference

gene sets are available for download on the nGASP wiki http://www.wormbase.org/wiki/index.php/nGASP and on the nGASP ftp site ftp://ftp.wormbase.org/pub/wormbase/nGASP.

The submitted gene predictions and the reference gene sets can be viewed in a genome browser based on GBrowse [22] at http://dev.wormbase.org/ngasp/. Each gene set is displayed in a different colour (Figure 3).

Authors' contributions

TF, LS, AC, PF, SM and DB participated in the design of the study, and LS oversaw the study. TF and LS organised the competition and liaised with the groups that submitted gene sets to nGASP. The nGASP consortium submitted gene sets to nGASP. SM processed the raw submitted gene sets to ensure they were in standard GFF3 format. TH set up the ftp site, and LS and SM set up the genome browsers to display the nGASP gene sets. AC analysed the accuracy of the submitted gene sets, identified features correlated with gene-finders' accuracies, and drafted the manuscript. PF also analysed the accuracy of the submitted gene sets. DB and SM used some of the gene-finders that were judged most accurate by nGASP to produce new gene sets for the C. elegans, C. briggsae, C. brenneri, C. remanei, C. japonica and Brugia malayi genomes. Two groups in the nGASP Consortium (those of M. Stanke and G. Rätsch) made gene sets for the six nematode genomes using their gene-finders. SM set up an ftp site for gene prediction resources and the final nGASP gene sets for these species. All authors read and approved the final manuscript.

The nGASP Consortium

GESECA: Darin Blasiar[1]

N-SCAN: Randall H. Brown[2], Michael R. Brent[2]

GENOMIX: Avril Coghlan[3], Richard Durbin[3]

GeneID and SGP2: Tyler Alioto[4], Francisco Câmara[4], Roderic Guigó[4]

SNAP: Ian Korf[5]

Gramene: Chengzhi Liang[6], Doreen Ware[6,7], Lincoln Stein[7]

GeneMark.hmm: Alex Lomsadze[8], Mark Borodovsky[8]

Agene: Kasper Munch[9], Anders Krogh[25]

Evigan and CRAIG: Qian Liu[10], Axel E. Bernal[10], Fernando C. N. Pereira[10], Aaron J. Mackey[11], David S. Roos[11]

MGENE: Gabriele Schweikert[12,13,14], Georg Zeller[12,14], Alexander Zien[12,15], Jonas Behr[12], Cheng Soon Ong[12,13], Petra Philips[12], Anja Bohlen[12], Regina Bohnert[12], Fabio De Bona[12], Sören Sonnenburg[15], Gunnar Rätsch[12]

GLEAN: Aaron Mackey[11], Qian Liu[10], Fernando C. N. Pereira[10], David S. Roos[11]

JIGSAW: Johnathan E. Allen[26], Steven Salzberg[16]

GlimmerHMM: Mihaela Pertea[16], Steven Salzberg[16]

EUGENE: Jérôme Gouzy[17], Céline Noirot[18], Thomas Schiex[18]

Fgenesh, Fgenesh++, Fgenesh++C: Peter Kosarev[19], Igor Seledsov[19], Vladimir Molodtsov[19], Victor Solovyev[19,20]

AUGUSTUS: Mario Stanke[21]

ExonHunter: Broňa Brejová[22], Tomás Vinar[22]

MAKER: Brandi L. Cantarel[23], Ian Korf[5], Sofia M.C. Robb[24], Genis Parra[5], Eric Ross[24], Barry Moore[23], Carson Holt[23], Alejandro Sánchez Alvarado[24], and Mark Yandell[23]

[1]Washington University School of Medicine, St Louis, MO 63108, USA

[2]Laboratory for Computational Genomics, Washington University, Campus Box 8510, 4444 Forest Park Ave, St. Louis, Missouri 63108, USA

[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom

[4]Bioinformatics and Genomics Program, Center for Genomic Regulation, Doctor Aiguader 88, E-08003, Barcelona, Spain

[5]Department of Molecular and Cellular Biology, University of California Davis, Davis, CA 95616, USA

[6]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, NY, 11724, USA

[7]USDA-ARS NAA Plant, Soil & Nutrition Laboratory Research Unit, Cornell University, Ithaca, NY, 14853, USA

[8]School of Biology, George Institute of Technology, Atlanta, Georgia, USA

[9]Centre for Comparative Genomics, Department of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark

[10]Computer and Information Science Department. University of Pennsylvania, USA

[11]Department of Biology, Penn Genomics Institute. University of Pennsylvania, USA

[12]Friedrich Miescher Laboratory, Max Planck Society, Spemannstr. 39, 72076 Tübingen, Germany

[13]Max Planck Institute for Biological Cybernetics, Spemanstr. 38, 72076 Tübingen, Germany

[14]Max Planck Institute for Developmental Biology, Spemanstr. 35, 72076 Tübingen,

[15]Fraunhofer Institute FIRST.IDA, Kekulestr. 7, 12489 Berlin, Germany

[16]Center for Bioinformatics and Computational Biology, Biomolecular Sciences Building, University of Maryland, College Park, MD 20742, USA

[17]Unité de Biométrie et Intelligence Artificielle, INRA, UR 875, BP 52627, Chemin de Borde Rouge, 31326, Auzeville, France

[18]Laboratoire Interactions Plantes Micro-organismes UMR441/2594, INRA/CNRS, F-31Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY,10549,USA

[19]Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY,10549,USA

[20]Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, United Kingdom

[21]Center for Biomolecular Science and Engineering, University of California Santa Cruz, USA

[22]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA

[23]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, 15 North 2030 East, Salt Lake City, UT 84112-5330, USA

[24]Department of Neurobiology and Anatomy, Howard Hughes Medical Institute, University of Utah School of Medicine, Salt Lake City, UT 84132, USA

[25]Bioinformatics Centre, Department of Molecular Biology, University of Copenhagen, Ole Maaloes Vej 5, 2200 Copenhagen N, Denmark

[26]Lawrence Livermore National Laboratory

Acknowledgements

We thank Michael Han (WormBase) for kindly providing us with the MLAGAN multi-genome alignments, John Spieth for useful discussions, and Payan Canaran for his GFF3 format checker. This work was supported in part by a grant from the National Human Genome Research Institute at the US National Institutes of Health # P41 HG02223. A. Coghlan is supported by an EMBO Long-Term Fellowship and the Wellcome Trust. P. Flicek is supported by EMBL. We particularly thank nGASP Consortium participants Gunnar Rätsch, Gabriele Schweikert and Mario Stanke for making new gene sets for the *Caenorhabditis* species and *Brugia malayi*.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Figures

Figure 1 - Accuracy of the Submitted Gene Sets.

Plots of the specificity against sensitivity of the submitted gene sets, at the base level (A), exon level (B), isoform level (C) and gene level (D). The submitted gene sets are coloured by nGASP category, with *ab initio* (category 1) gene sets in red, gene-finders that used multi-genome alignments (category 2) in black, gene-finders that used transcript/protein alignments (category 3) in blue, and combiners (category 4) in green. The gene sets are labelled as follows: AU: AUGUSTUS, MG: MGENE, CR: CRAIG, AG: Agene, EU: EUGENE, FPC: Fgenesh++C, FP: Fgenesh++, FG: Fgenesh, GE: GeneID, GM: GeneMark.hmm, GX: GENOMIX, GS: GESECA, GN: GLEAN, GL:

GlimmerHMM, GR: Gramene, JW: JIGSAW, MK: MAKER, MG: MGENE, NS: N-SCAN, SG: SGP2, SN: SNAP, EX: ExonHunter, EV: Evigan.

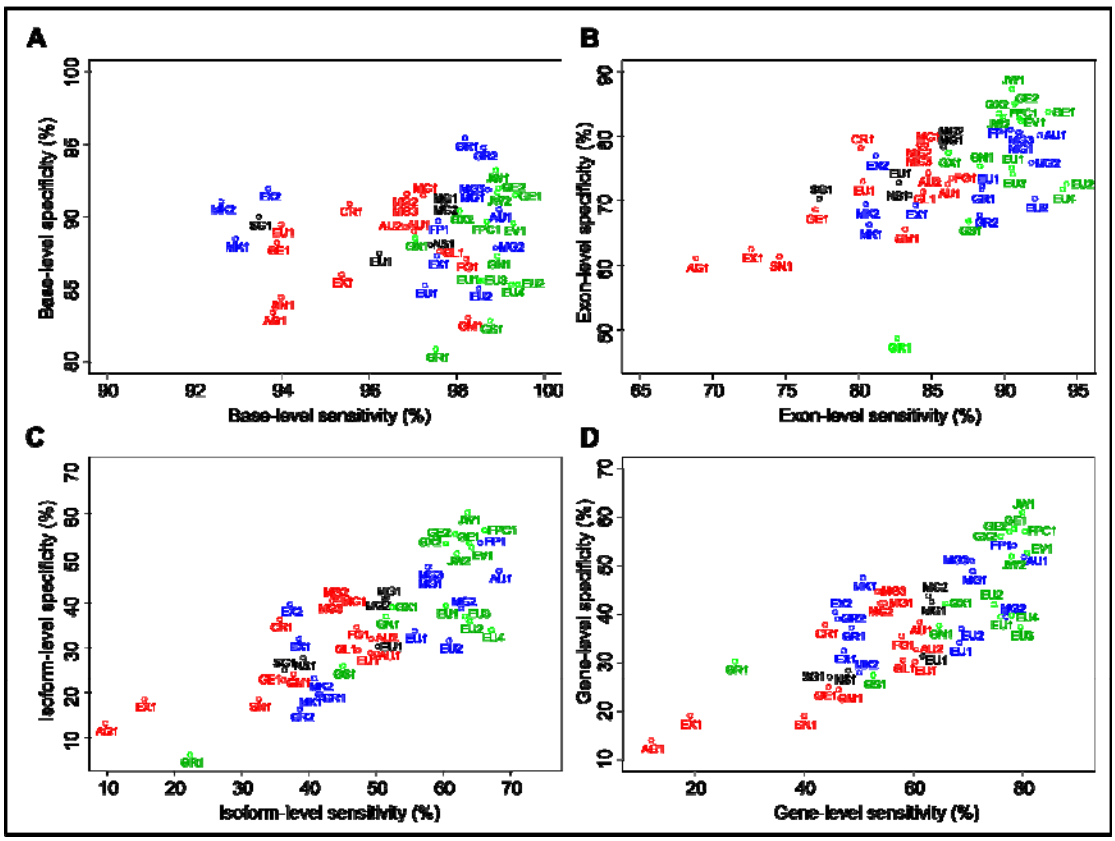


Figure 2 – Factors Affecting Gene-finding Accuracy.

Plots of gene-level sensitivity against features of genes that are correlated with gene-finding accuracy: (A) the lowest hexamer score of any of the exons in the gene, (B) the number of exons in the gene, (C) the length of the shortest exon in the gene, (D) the length of the longest intron in the gene, (E) the strength of the translation start signal, (F) the lowest score of any of splice sites in the gene, (G) the percent identity with the *C. briggsae* ortholog at the amino acid level, (H) the maximum distance to a neighbouring gene, and (I) the number of isoforms in the gene. In each plot, the

submitted gene sets are coloured by nGASP category, with *ab initio* (category 1) gene sets in red, gene-finders that used multi-genome alignments (category 2) in black, and gene-finders that used transcript/protein alignments (category 3) in blue. The solid lines show the median sensitivities of the gene sets in a category, while the dashed lines show the maximum sensitivity of the gene sets in a category.

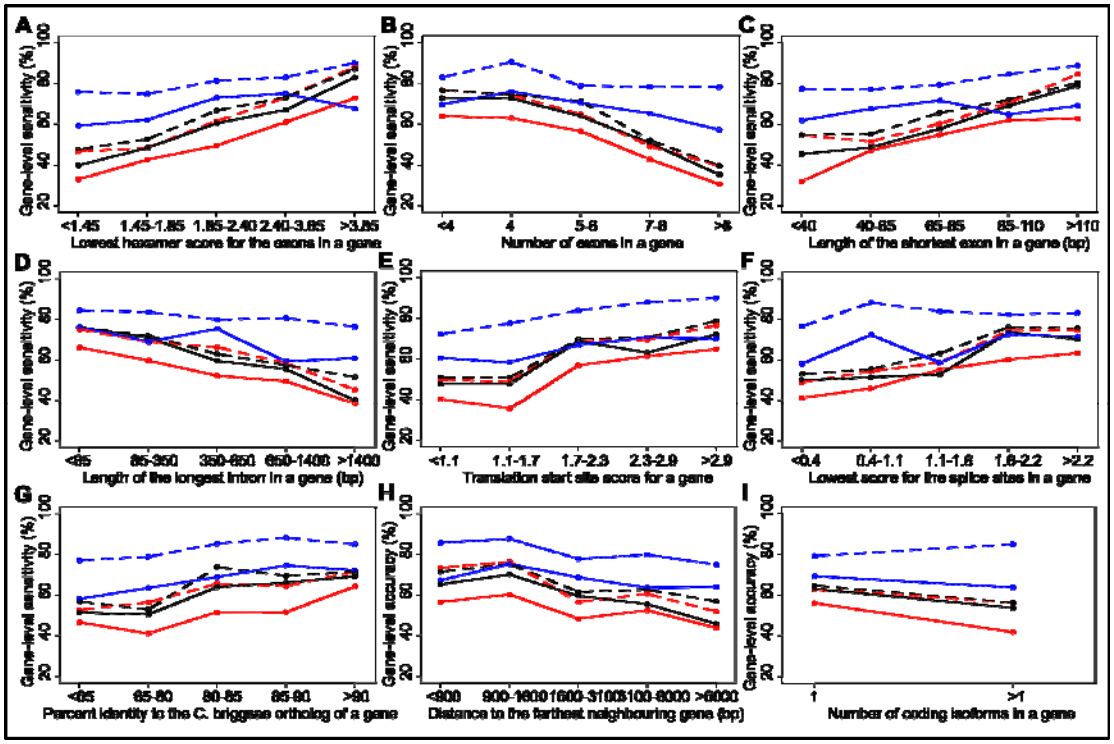
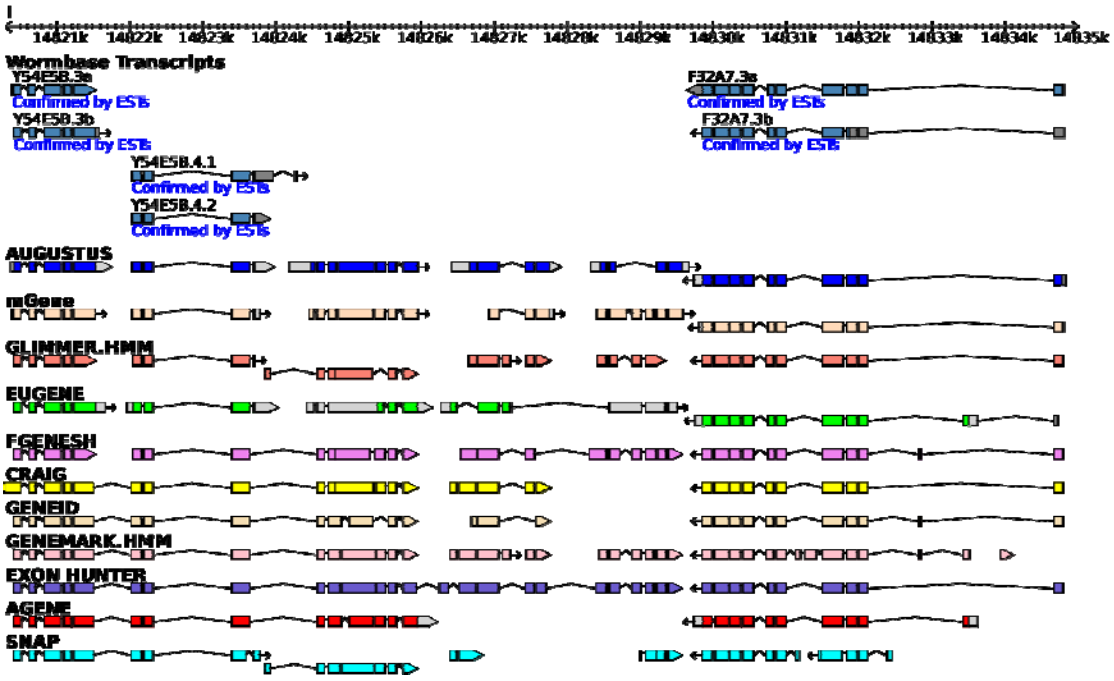


Figure 3 - A Screenshot from the nGASP Genome Browser.

This shows part of an nGASP test region on chromosome I, with the curated WormBase gene models and the *ab initio* (category 1) gene sets submitted to nGASP for that region.



Tables

Table 1 - Participating Groups and Submitted Gene Sets.

The research groups that participated in nGASP, the names of the software used to produce gene prediction sets, and the number of gene sets submitted in each of the nGASP categories by a research group, are given. Here 'cat3:2' means that 2 gene sets in category 3 were submitted. In some cases a group submitted two gene sets produced by using different parameters of their software to the same nGASP category.

Participating group	Program name	Number of gene sets submitted in each category
Blasiar et al, Saint Louis, USA	GESECA (D. Blasiar, unpublished)	cat4:1
Borodovsky et al, Atlanta, USA	GeneMark.hmm[23]	cat1:1
Brent et al, Saint Louis, USA	N-SCAN [24]	cat2:1
Durbin et al, Cambridge, UK	GENOMIX [25]	cat4:2
Guigó et al, Barcelona, Spain	GeneID[1] [26], SGP2 [27]	GeneID: cat1:1, cat4:2; SGP2: cat2:1

Korf et al, Davis, USA	SNAP [28]	cat1:1
Krogh et al, Copenhagen, Denmark	Agene [29]	cat1:1
Liang et al, Cold Spring Harbor, USA	Gramene (Liang et al, unpublished)	cat3:2, cat4:1
Pereira et al, Pennsylvania, USA	Evigan (Q. Liu et al, manuscript in prep.), CRAIG [30]	CRAIG: cat1:1; Evigan: cat4:1
Rätsch et al, Tübingen, Germany	MGENE (Schweikert et al, submitted)	cat1:3, cat2:2, cat3:3
Roos et al, Pennsylvania, USA	GLEAN [31]	cat4:1
Salzberg et al, Maryland, USA	JIGSAW [13], GlimmerHMM [13]	GlimmerHMM: cat1:1; JIGSAW: cat4:2
Schiex et al, Toulouse, France	EUGENE [32]	cat1:1, cat2:1, cat3:2, cat4:4
Solovyev et al, University of London and Softberry Inc, New York, USA	Fgenesh, Fgenesh++, Fgenesh++C [12]	Fgenesh: cat1:1; Fgenesh++: cat3:1; Fgenesh++C: cat4:1
Stanke et al, Santa Cruz, USA	AUGUSTUS [11]	cat1:2, cat3:1
Vinar et al, New York, USA	ExonHunter [33]	cat1:1, cat3:2
Yandell et al, Berkeley, USA	MAKER [34]	cat3:2

[1]The GeneID gene set was submitted after the nGASP deadline.

Table 2 - The nGASP Test and Training Genomic Regions.

The ten 1-Mb regions of the *C. elegans* genome provided to the nGASP participants for training their gene-finders, and ten 1-Mb test regions in which they were asked to make gene predictions for the nGASP assessment.

Type of nGASP region	**Criterion used for selecting region**	**Coordinates in the *C. elegans* WS160 genome**
Training	High conservation, high gene density, autosomal	II: 2000001-3000000
Training	High conservation, high gene density, autosomal	V: 9000001-10000000
Training	High conservation, low gene density, autosomal	III: 1000001-2000000
Training	High conservation, low gene density, autosomal	IV: 2000001-3000000
Training	Low conservation, high gene density, autosomal	I: 12000001-13000000
Training	Low conservation, high gene density, autosomal	V: 4000001-5000000
Training	Low conservation, low gene density, autosomal	I: 2000001-3000000
Training	Low conservation, low gene density, autosomal	II: 13000001-14000000
Training	High conservation, low gene density, X-chromosome	X: 3000001-4000000
Training	High conservation, low gene density, X-chromosome	X: 2000001-3000000
Test	High conservation, high gene density, autosomal	IV: 7000001-8000000
Test	High conservation, high gene density, autosomal	V: 12000001-13000000
Test	High conservation, low gene density, autosomal	IV: 1-1000000
Test	High conservation, low gene density, autosomal	I: 14000001-15000000
Test	Low conservation, high gene density, autosomal	V: 16000001-17000000
Test	Low conservation, high gene density, autosomal	II: 1-1000000
Test	Low conservation, low gene density, autosomal	IV: 14000001-15000000
Test	Low conservation, low gene density, autosomal	I: 1000001-2000000
Test	High conservation, low gene density, X-chromosome	X: 4000001-5000000
Test	High conservation, low gene density, X-chromosome	X: 8000001-9000000

Table 3 – Evaluation of Submitted Gene Sets.

The accuracy of the submitted gene sets evaluated using the reference gene sets **ref1** and **ref2**. The sensitivity (Sn) results are given for reference set **ref1**, and the specificity (Sp) results are given for set **ref2**. The gene sets are divided according to nGASP category, where category 1 is *ab initio* gene-finders, 2 is gene-finders that used multi-genome alignments, 3 is gene-finders that used alignments of ESTs, mRNAS and proteins, and 4 is combiners.

Gene set	**nGASP category**	**Base Sn, ref1**	**Base Sp, ref2**	**Exon Sn, ref1**	**Exon Sp, ref2**	**Isoform Sn, ref1**	**Isoform Sp, ref2**	**Gene Sn, ref1**	**Gene Sp, ref2**
Agene	1	93.79	83.41	68.87	61.09	9.75	13.12	11.97	14.08
AUGUSTUS v1	1	97.02	89.02	86.12	72.55	50.08	28.65	61.05	38.41
AUGUSTUS v2	1	96.82	89.30	84.77	74.33	49.26	31.92	60.45	32.74
CRAIG	1	95.55	90.91	80.17	78.15	35.70	36.30	43.81	37.80
EUGENE	1	93.98	89.48	80.28	73.00	49.09	28.82	60.24	30.17
ExonHunter	1	95.36	86.03	72.63	62.53	15.54	18.58	19.07	19.18
Fgenesh	1	98.21	87.11	86.37	73.55	47.11	34.59	57.81	35.43
GeneID	1	93.89	88.24	77.04	68.63	36.20	22.84	44.42	25.08
GeneMark.hmm	1	98.26	83.06	83.17	65.58	37.69	23.98	46.25	24.54
GlimmerHMM	1	97.60	87.62	84.42	71.37	47.27	29.31	58.01	30.55
MGENE v1	1	97.23	91.48	84.63	78.58	44.63	40.88	54.77	42.29
MGENE v2	1	96.86	91.60	84.17	78.70	43.97	40.87	53.96	42.35
MGENE v3	1	96.86	91.59	84.17	78.63	43.47	40.50	53.35	44.75
SNAP	1	93.98	84.47	74.57	61.30	32.56	18.61	39.96	19.09
EUGENE	2	96.23	87.48	82.75	72.82	50.25	30.19	61.66	31.36
MGENE v1	2	97.70	90.91	85.81	78.35	51.57	41.18	63.29	42.52

MGENE v2	2	97.70	90.91	85.81	78.30	51.24	40.87	62.68	43.83
N-SCAN	2	97.39	88.07	83.51	70.83	39.17	27.69	48.07	28.39
SGP2	2	93.47	89.99	77.32	70.27	36.36	24.89	44.62	27.11
AUGUSTUS v1	3	98.96	90.52	92.45	80.20	68.26	47.07	80.12	51.81
EUGENE v1	3	97.27	85.30	88.49	72.22	55.70	33.70	68.36	34.15
EUGENE v2	3	98.50	85.09	92.10	70.32	60.83	31.53	68.76	36.100
ExonHunter v1	3	97.55	87.31	83.90	69.33	38.51	31.92	47.26	32.52
ExonHunter v2	3	93.69	91.96	81.18	76.92	37.19	39.74	45.64	40.47
Fgenesh++	3	97.57	89.70	90.43	80.93	65.45	53.44	78.30	54.20
Gramene v1	3	98.19	95.42	88.45	71.76	41.65	19.55	48.68	37.20
Gramene v2	3	98.61	94.77	88.31	67.77	38.68	16.27	46.04	39.04
MAKER v1	3	92.94	88.50	80.73	66.27	41.32	19.62	50.71	47.55
MAKER v2	3	92.61	91.05	80.49	69.49	40.83	23.19	50.10	27.95
MGENE v1	3	98.70	91.89	90.96	80.67	57.69	48.04	70.79	48.89
MGENE v2	3	98.88	87.86	91.86	75.88	62.64	38.72	76.88	39.46
MGENE v3	3	98.71	91.88	90.99	80.61	57.69	48.00	70.59	51.11
EUGENE v1	4	98.53	85.60	90.50	75.07	60.38	39.29	75.90	39.51
EUGENE v2	4	99.40	85.35	94.27	72.63	63.90	35.89	74.70	42.04
EUGENE v3	4	98.57	85.60	90.57	74.18	63.26	36.91	79.52	37.36
EUGENE v4	4	99.24	85.34	93.99	71.76	67.09	33.94	77.91	39.76
Evigan	4	99.29	89.59	91.13	82.31	64.22	52.38	80.72	52.71
Fgenesh++C	4	98.69	89.68	91.06	82.70	66.13	56.26	80.32	57.14
GeneID v1	4	99.34	91.50	93.01	83.78	63.90	53.27	78.31	57.67
GeneID v2	4	98.96	91.97	90.71	85.03	61.66	55.49	77.51	57.10
GENOMIX v1	4	97.05	88.55	86.16	77.36	52.40	39.04	65.86	42.21
GENOMIX v2	4	98.07	90.40	89.66	83.53	60.38	53.27	75.90	56.05
GESECA	4	98.76	82.84	87.56	66.81	45.05	25.86	52.61	27.43
GLEAN	4	98.92	87.28	88.33	75.37	51.44	36.95	64.66	37.62
Gramene	4	97.52	80.89	82.67	48.71	22.36	6.09	27.31	30.32

JIGSAW v1	4	98.89	93.22	90.50	87.36	63.58	60.23	79.92	61.00
JIGSAW v2	4	98.92	91.66	89.94	83.04	61.98	51.07	77.91	51.95

References

1. The *C. elegans* Sequencing Consortium: Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998, 282(5396):2012-2018.

2. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A *et al*: The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol* 2003, 1(2):E45.

3. Sternberg PW, Waterston RH, Speith J, Eddy SR, Wilson RK: Genome sequence of additional *Caenorhabditis* species: enhancing the utility of *C. elegans* as a model organism. In.: National Human Genome Research Institute; 2003.

4. Rogers A, Antoshechkin I, Bieri T, Blasiar D, Bastiani C, Canaran P, Chan J, Chen WJ, Davis P, Fernandes J *et al*: WormBase 2007. *Nucleic Acids Res* 2008, 36(Database issue):D612-617.

5. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D *et al*: Draft genome of the filarial nematode parasite Brugia malayi. *Science* 2007, 317(5845):1756-1760.

6. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A: Critical assessment of methods of protein structure prediction (CASP)--round 6. *Proteins* 2005, 61 Suppl 7:3-7.

7. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: Genome annotation assessment in Drosophila melanogaster. *Genome Res* 2000, 10(4):483-501.

8. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E *et al*: EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol* 2006, 7 Suppl 1:S2 1-31.

9. Durbin R, Thierry-Mieg J: The ACeDB Genome Database. In: *Computational Methods in Genome Research.* Edited by Suhai S. New York: Plenum Press; 1994: 45-56.

10. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L *et al*: TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* 2006, 34(Database issue):D572-580.

11. Stanke M, Schoffmann O, Morgenstern B, Waack S: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 2006, 7:62.

12. Salamov AA, Solovyev VV: *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* 2000, 10(4):516-522.

13. Allen JE, Majoros WH, Pertea M, Salzberg SL: JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. *Genome Biol* 2006, 7 Suppl 1:S9 1-13.

14. Kent WJ, Zahler AM: Conservation, regulation, synteny, and introns in a large-scale *C. briggsae-C. elegans* genomic alignment. *Genome Res* 2000, 10(8):1115-1125.

15. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 2003, 13(4):721-731.

16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402.

17. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: FlyBase: genomes by the dozen. *Nucleic Acids Res* 2007, 35(Database issue):D486-491.

18. Nash R, Weng S, Hitz B, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE *et al*: Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res* 2007, 35(Database issue):D468-471.

19. Consortium U: The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2007, 35(Database issue):D193-197.

20. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: Ensembl 2007. *Nucleic Acids Res* 2007, 35(Database issue):D610-617.

21. Pruitt KD, Tatusova T, Maglott DR: NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007, 35(Database issue):D61-65.

22. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A *et al*: The generic genome browser: a building block for a model organism system database. *Genome Res* 2002, 12(10):1599-1610.

23. Besemer J, Lomsadze A, Borodovsky M: GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001, 29(12):2607-2618.

24. Gross SS, Brent MR: Using multiple alignments to improve gene prediction. *J Comput Biol* 2006, 13(2):379-393.

25. Coghlan A, Durbin R: Genomix: a method for combining gene-finders' predictions, which uses evolutionary conservation of sequence and intron-exon structure. *Bioinformatics* 2007, 23(12):1468-1475.

26. Parra G, Blanco E, Guigo R: GeneID in Drosophila. *Genome Res* 2000, 10(4):511-515.

27. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigo R: Comparative gene prediction in human and mouse. *Genome Res* 2003, 13(1):108-117.

28. Korf I: Gene finding in novel genomes. *BMC Bioinformatics* 2004, 5:59.

29. Munch K, Krogh A: Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics* 2006, 7:263.

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

31. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: Creating a honey bee consensus gene set. *Genome Biol* 2007, 8(1):R13.

32. Foissac S, Schiex T: Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics* 2005, 6:25.

33. Brejová B, Brown DG, Li M, Vinar T: ExonHunter: a comprehensive approach to gene finding. *Bioinformatics* 2005, 21 Suppl 1:i57-65.

34. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M: MAKER: An Easy-to-use Annotation Pipeline Designed for Emerging Model Organism Genomes. *Genome Res* 2007, 18:188.