BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 606–613
doi:10.1093/bioinformatics/btt577

Genome analysis

Advance Access publication October 11, 2013

GIIRA—RNA-Seq driven gene finding incorporating ambiguous reads

Franziska Zickmann, Martin S. Lindner and Bernhard Y. Renard*

Research Group Bioinformatics (NG4), Robert Koch-Institute, Nordufer 20, 13353 Berlin, Germany

Associate Editor: John Hancock

ABSTRACT

Motivation: The reliable identification of genes is a major challenge in genome research, as further analysis depends on the correctness of this initial step. With high-throughput RNA-Seq data reflecting currently expressed genes, a particularly meaningful source of information has become commonly available for gene finding. However, practical application in automated gene identification is still not the standard case. A particular challenge in including RNA-Seq data is the difficult handling of ambiguously mapped reads.

Results: We present GIIRA (Gene Identification Incorporating RNA-Seq data and Ambiguous reads), a novel prokaryotic and eukaryotic gene finder that is exclusively based on a RNA-Seq mapping and inherently includes ambiguously mapped reads. GIIRA extracts candidate regions supported by a sufficient number of mappings and reassigns ambiguous reads to their most likely origin using a maximum-flow approach. This avoids the exclusion of genes that are predominantly supported by ambiguous mappings. Evaluation on simulated and real data and comparison with existing methods incorporating RNA-Seq information highlight the accuracy of GIIRA in identifying the expressed genes.

Availability and implementation: GIIRA is implemented in Java and is available from https://sourceforge.net/projects/giira/.

Contact: renardB@rki.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 8, 2013; revised on September 26, 2013; accepted on September 27, 2013

1 INTRODUCTION

The reliable annotation of genes as the regions encoding the basis for all processes in the cell is a key goal of genomic research. Hence, numerous studies focus on revealing the structure of genes and their controlling mechanisms to enhance the understanding of the functionality of proteins and their interactions (Schrimpe-Rutledge *et al*., 2012; Wang *et al*., 2012; Wijaya *et al*., 2013). Following the need for accurate gene prediction methods, various gene finders have been developed identifying genes either by genome sequence comparison, *ab initio* by using the genomic sequence alone, evidence based by integrating different kinds of additional external information (Goodswen *et al*., 2012).

Ab initio gene finders predict genes based on the presence of open reading frames (ORFs) in the genome and in case of eukaryotes identify intron-exon structures indicated by known splice sites (Delcher *et al*., 2007; Korf, 2004; Lukashin and Borodovsky, 1998; Majoros *et al*., 2004). Typically, these approaches are based on statistical or machine learning techniques such as Hidden Markov Models, and they require training data to evaluate the probability for each gene and gene structure (Goodswen *et al*., 2012).

In contrast to *ab initio* methods, evidence-based and comparative gene finders make use of additional external information to identify genes and their structures. Sources of information include EST libraries, messenger RNA or protein sequences. The external evidence is compared with the genome of interest to identify regions showing similarity to the given sequences (Allen and Salzberg, 2005; Savidor *et al*., 2006; Wei and Brent, 2006). Hybrid approaches, such as AUGUSTUS (Stanke *et al*., 2006), combine *ab initio* gene prediction with evidence from other sources to verify the predicted genes. For an overview on gene finding algorithms, the reader is referred to Goodswen *et al*. (2012) or Guigó *et al*. (2006).

Despite all efforts, gene identification still faces significant challenges handling complex gene structures, rare splice sites or mutations in genes (Ederveen *et al*., 2013; Goodswen *et al*., 2012). These problems can be approached by using the knowledge available from high-throughput RNA-Seq experiments (Wang *et al*., 2009). The transcriptome reflects the genes expressed in the current condition of the cell, which provides valuable information to identify novel genes or to confirm predicted genes. Although RNA-Seq experiments were included in several annotation studies (Martin *et al*., 2010; Palmieri *et al*., 2012; Pickrell *et al*., 2012; Sultan *et al*., 2008; Tu *et al*., 2012), so far only few gene finders directly incorporate RNA-Seq in gene prediction.

Methods for gene expression analysis such as iReckon (Mezlini *et al*., 2013), Cufflinks (Trapnell *et al*., 2010) and Erange (Mortazavi *et al*., 2008) perform a transcript assembly on RNA-Seq reads and thereby allow the identification of exons and splice sites, but they do not predict reading frames and start and stop codon for genes. For an overview of transcriptome annotation, the reader is referred to Garber *et al*. (2011). AUGUSTUS allows the integration of RNA-Seq experiments as an additional external source for eukaryotic gene identification (Stanke *et al*., 2008), whereas GeneMark (Besemer *et al*., 2001; Martin *et al*., 2010) incorporates RNA-Seq evidence on prokaryotic gene predictions to identify operons. The gene finder G-Mo.R-Se (Denoeud *et al*., 2008) predicts gene models based on RNA-Seq reads, but does not identify mono-exonic genes and only incorporates non-ambiguous mappings.

Because for instance repetitive or highly similar regions, or homologous genes lead to a substantial part of non-unique

*To whom correspondence should be addressed.

mappings, discarding ambiguously mapped reads from further analysis may result in a significant loss of prediction accuracy.

To use the complete information contained in RNA-Seq experiments for gene identification, we developed a RNA-Seq-based *de novo* gene predictor called GIIRA (**G**ene **I**dentification **I**ncorporating **R**NA-Seq and **A**mbiguous reads) that works on a reference genome and reads derived in a RNA-Seq experiment. GIIRA is primarily focused on prokaryotic gene prediction and in particular resolves genes within the continuously expressed region of an operon. However, GIIRA can also be applied to predict genes and alternative transcripts for eukaryotes, and it leverages information from spliced reads for intron identification. Hence, it is also a useful addition to annotation pipelines, such as MAKER (Holt and Yandell, 2011), or a good complement to other eukaryotic gene finders. The identified transcripts are completed into gene models via a search for start and stop codons as well as reading frame and strand prediction. Based on the observed mapping coverage, GIIRA identifies candidate genes that are refined in further validating steps. Ambiguous reads are reassigned to their most likely origins using a maximum-flow approach formulated as a linear program.

In contrast to other approaches to ambiguous read assignment, such as the expectation maximization-based strategy introduced in Chung *et al*. (2011) or ContextMap (Bonfert *et al*., 2012), our approach can integrate information on the likelihood of a read alignment not only from a fixed context (interval of specified length) or a context solely based on the mapping. Instead, we directly incorporate the information gained in the process of identifying gene candidates and further the linear program ensures a convergence to an optimal solution.

In principle, the general idea of the maximum-flow approach can also be applied to other questions related to ambiguity resolving.

We validate the accuracy of GIIRA in three simulations and compare our approach with the widely used method Cufflinks as well as the gene finders GeneMark, GLIMMER3 (Delcher *et al*., 2007) and AUGUSTUS. Finally, we apply GIIRA to two real datasets including ~11 million reads from an *Escherichia coli* and ~6 million reads from a *Saccharomyces cerevisiae* RNA-Seq experiment.

2 METHODS

As depicted in Figure 1, the proposed algorithm consists of four steps. The input of GIIRA is a set of RNA-Seq reads that are aligned to a reference genome using an external alignment method (Fig. 1A). Based on the alignment, GIIRA identifies regions on the genome that are likely to be expressed genes, in the following called *gene candidates* (Fig. 1B). The identification regards the nucleotide coverage as well as splicing events indicated by the RNA-Seq reads. For prokaryotes, these candidates are regarded as expressed regions that might contain more than one gene, hence they are refined to determine the correct gene structure. Finally, ambiguously mapped reads are reallocated to their most likely origins using a maximum-flow optimization approach (Fig. 1C). Based on this reassignment, the candidate genes undergo a refinement leading to the erasing of candidate genes and isoforms without a sufficient number of remaining supporting reads (Fig. 1D).

2.1 Alignment analysis

GIIRA is based on an alignment of reads from a RNA-Seq experiment to the DNA sequence of interest. For eukaryotes it is advisable—although not strictly necessary—to use a split read mapper for this alignment to obtain support for splicing events. GIIRA is preconfigured to call either TopHat2 (Kim *et al*., 2013) or BWA (Li and Durbin, 2009) for read mapping, but can include the results of any read mapper with output in SAM format (Li *et al*., 2009). GIIRA takes all mappings reported in the resulting SAM file into account, which includes one mapping for unique reads and several for ambiguous reads. For performance reasons, we only store the start positions of reads and their differences to the reference, as well as read quality and potential splice sites.

2.2 Candidate search

2.2.1 Extraction As illustrated in Figure 1B, regions with sufficient support of mapped reads are extracted to serve as *candidate genes*. The algorithm traverses all start positions of read alignments and tests if the coverage at these positions exceeds a minimum coverage. If this is the case, a new candidate gene is opened and all following reads are assigned to the currently open region. This process is continued until the coverage falls below the coverage threshold. Then the current candidate gene is closed and we search for a new region exceeding the minimum coverage. As the coverage threshold is a crucial parameter in the analysis, it can either be estimated from the given data without any *a priori* knowledge or be defined by the user.

In case of splicing events, this basic procedure is extended: A splice site is only considered as a non-erroneous site if it has a sufficient support of reads. By default the threshold for splice site acceptance is set equal to the overall desired minimum coverage. In case reads overlap an accepted splice site, they are assigned to their corresponding isoform, e.g. an intron starting at this splice position or an ongoing exon. During this first candidate extraction all isoforms with sufficient support by reads are taken into account, the refinement and exclusion of erroneous alternative isoforms are performed in subsequent steps (see Section 2.4).

Details on the candidate search and the choice of coverage thresholds are given in the Supplementary Material.

2.2.2 Prokaryotic gene structuring Prokaryotic candidates undergo an additional extraction step, as prokaryotic operons contain a continuously expressed region including one or more genes that have to be identified respecting the present ORFs. To determine the most likely

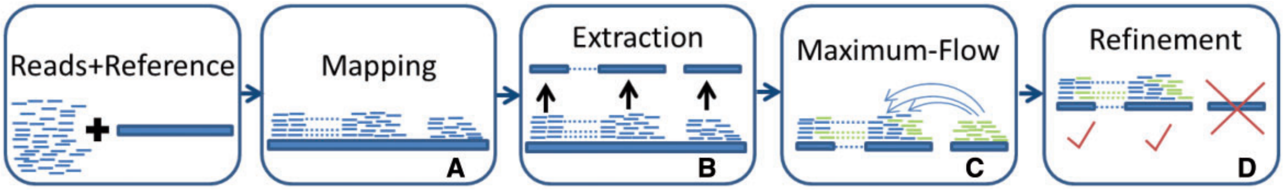

Fig. 1. Workflow of GIIRA: given a genomic sequence and a set of RNA-Seq reads, reads are mapped to the reference (**A**) and the resulting alignment is then analyzed by GIIRA. Candidate genes are extracted (**B**) and ambiguous reads are reassigned using a maximum-flow optimization (**C**). Finally, candidate genes are evaluated based on the reallocated reads (**D**)

gene structure, first all forward and reverse ORFs of the candidate sequence are extracted. Second, the direction is selected that provides a set of ORFs that covers a large number of bases in this operon while restricting the overall number of ORFs. To achieve a trade-off between these two goals, we adopt and alter a scoring metric from alignment evaluations (Vingron and Waterman, 1994).

The set of all possible ORFs in a candidate sequence with length L is denoted as O. An ORF $o_i \in O$ contributes with its length l_i to the number of covered bases; hence it is assigned a positive ('match') score $m_i = l_i$. If two or more ORFs o_i and o_j overlap, the overlapping region is assigned a negative score ov_{ij} such that no region is counted twice. To avoid the suboptimal solution of simply selecting all ORFs present in O, we enforce sparsity by introducing an *ORF open penalty* p_i for each ORF o_i:

$$p_i = -\left(\frac{L}{l_i} \cdot \frac{l_{max}}{l_i}\right),$$

with l_{max} denoting the length of the longest ORF included in O. This penalty is smaller for longer ORFs, as these are preferable to short ones because they cover more bases. Further, p_i reflects whether o_i is comparably short or long in relation with the ORFs present in O.

This can be combined in a linear program that maximizes the sum of all scores:

$$max \sum_{i \in O} (m_i + p_i) + \sum_{i \neq j} ov_{ij},$$

Details on the scoring metric and the linear program are given in the Supplementary Material.

2.3 Maximum-Flow optimization

In previous steps, all read mappings contributed equally to the extraction of candidate regions, even if a read had multiple mappings with similar quality.

However, as each read can only arise from one genomic locus, we aim at reassigning ambiguously mapped reads to their most likely origin. To do so, we use a maximum-flow representation depending on the gathered information of extracted gene candidates.

The rationale behind this approach is that if several genes compete for the same read, their overall read coverage and the presence of support from unique reads indicate the most likely origin of this read. Both factors do not only enhance the probability for a candidate to be chosen, but also decrease the chances of the competitors such that the number and quality of the competitors directly affect the choice for the best origin. Further, also the ambiguity of the read itself is taken into account by weighting the influence of reads on candidate quality by the number of their alignment positions. The more alignments a read has, the less it supports each single gene it is mapped to.

The problem of assigning each read to exactly one gene candidate can be formulated as a network problem as illustrated in Figure 2. We define a network $G = \{N, E\}$ with edge set E and node set $N = R \cup C \cup s \cup t$ with nodes $r \in R$ representing reads and nodes $c \in C$ representing gene candidates, respectively. Source node s and target node t are defined for technical reasons. Further, all edges are directed and an edge $e_{ij} \in E$ between two nodes represents that read $r_i \in R$ is assigned to gene $c_j \in C$. Note that each edge has a capacity, which can be understood as the maximal input that can pass through this edge. In contrast, nodes have an unlimited throughput.

The aim of the maximum-flow is to set all capacities φ_{ij} (belonging to edges e_{ij} connecting a read r_i to a candidate c_j) in a way that the flow passing from source to target node is maximized:

$$max \sum_{e_{ij} \in E} \varphi_{ij},$$

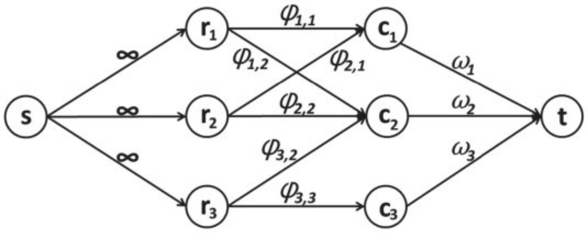

Fig. 2. Simplified example for a maximum-flow network representation passing flow from source node s to target node t. The source node is connected to the nodes representing reads (r_i), which are connected with all genes they were mapped to (c_j). The edge labels indicate the capacity for the throughput that is allowed to be passed from one node to the other (representing the support of the read to the corresponding candidate gene)

Each edge originating from the source has an unlimited capacity. The capacity φ_{ij} of the edges connecting reads and their possible corresponding genes is restricted by the following condition:

$$0 \leq \varphi_{ij} \leq y_{ij},$$

where $y_{ij} \in \{0, 1\}$ are the binary variables that denote whether the read r_i is assigned to gene c_j ($y_{ij} = 1$) or not ($y_{ij} = 0$). In other words, if a read is assigned to a gene, the corresponding edge connecting both nodes has a capacity with a maximal value of 1. If the read is not assigned, the capacity is zero.

In addition, we require all multiple reads to be assigned to exactly one candidate, as reflected in the constraint $\sum_j y_{ij} = 1$. Each gene has a maximal number of reads that can be assigned, depending on the support of reads for this gene and the support for its competitors. Because for each node the input flow has to equal the output flow, this maximum is given by the capacity ω_j of the edges connecting gene nodes to the target node:

$$\sum_{i | e_{ij} \in E} \varphi_{ij} \leq \omega_j,$$

where ω_j is calculated as follows:

$$\omega_j = \frac{b_j}{\sum\limits_{c_k \in P_j} b_k^u},$$

Here, b_j is the average base coverage of gene c_j derived by all its mapping reads, where in contrast b_j^u is the coverage derived only by reads that map uniquely to the corresponding gene. The set P_j contains all genes that directly compete with c_j for ambiguously mapped reads, or in other words, that share reads with gene c_j. For illustration, refer to Figure 2: here P_2 consists of c_1 and c_3, whereas P_1 only includes c_2 because c_1 only shares reads with c_2.

Allowing genes to influence their competitors with the help of their own likeliness ensures that genes with an overall high coverage are preferred over genes with less coverage. Otherwise genes with no or only few unique reads could be preferred over genes with a high unique coverage, as long as they have enough multiple hits.

The maximum-flow problem is formulated as an integer linear program including the constraints described earlier in the text. This program is solved using the IBM CPLEX academic version V12.4 (CPLEX, 2011) or, as a slower alternative, the open-source GLPK solver (GLPK, 2006).

2.4 Candidate refinement and scoring

The maximum-flow optimization identifies a unique position for each read such that the previously extracted gene candidates have to be refined according to the new assignment of reads. If a gene candidate or an alternative isoform lost all of its supporting reads, it is regarded as an artifact of ambiguous read mappings and is thus erased. All remaining genes are evaluated in a scoring process according to their exon length l_j, their read coverage and the quality of their assigned reads. It is also of relevance whether the corresponding reads are mapped ambiguously, as ambiguity implies more uncertainty for the gene and thus leads to a smaller score. The final gene score s_j for gene c_j is calculated as follows:

$$s_j = \frac{1}{l_j} \cdot \sum_{i|e_{ij} \in E} \frac{l_i \cdot q_i}{M_i},$$

where q_i denotes the quality of read, r_i, l_i its length and M_i its total number of mappings. GIIRA reports the identified genes and transcripts in GTF annotation format, including additional information on coverage and ambiguous read support. This allows an easy post-processing to verify genes for follow-up analyses (refer to the Supplementary Material for details).

3 EXPERIMENTAL SETUP

To evaluate GIIRA with regard to prediction accuracy and to compare it with existing methods given a known ground truth, we use a variety of different datasets to avoid any design bias toward a specific organism. We generated a prokaryotic simulated dataset based on *E.coli* (NCBI-Accession: NC_000913.2) and two eukaryotic simulations based on chromosome 15 of the human genome (NC_000015.9) and chromosome 4 of *S.cerevisiae* (NC_001136.10), respectively. Based on these data, we compare GIIRA with Cufflinks, GLIMMER3 and GeneMark in the prokaryotic simulation and to Cufflinks and AUGUSTUS in the eukaryotic simulations.

As GeneMark is originally an *ab initio* gene predictor that does not include RNA-Seq information, we used the framework proposed in Martin *et al*. (2010) that combines GeneMarkS (Besemer *et al*., 2001) *ab initio* predictions with the program ParseRnaSeq to include RNA-Seq evidence (refer to the Supplementary Material for details on the applied pipeline). Note that in this framework the resulting predictions cover operons rather than structural genes.

Further, to demonstrate the influence of ambiguous mappings on the prediction accuracy, we configured and compared a second version of GIIRA that excludes ambiguous mappings from further analysis.

The simulation setup uses the read simulator Mason (Holtgrewe, 2010) applied to the NCBI reference annotation for each organism of interest. In this annotation the coding sequence of each known isoform appears as a consecutive sequence; hence, the simulated reads show similar characteristics as real RNA-Seq reads because they cover alternative isoforms, span introns (if existing in the dataset) and show a coverage profile typical for gene expression. The simulated reads were aligned to the reference genome using TopHat2 (Kim *et al*., 2013), and the resulting alignment served as the starting point for all compared methods.

To demonstrate the performance of GIIRA on a real prokaryotic dataset, we applied GIIRA, Cufflinks, GLIMMER3 and GeneMark to a mapping of 11 million reads (NCBI-Accession: SRX180743) from *E.coli*. This dataset contains a large proportion of ambiguous mappings as well as high coverages in the areas coding for ribosomal RNA, posing a challenge to distinguish false from correct gene loci. As GIIRA is also applicable to eukaryotic organisms, a proof of principle application to a real *S.cerevisiae* dataset comprising 6 million reads (SRX187114) was performed comparing GIIRA and Cufflinks. For detailed information on the experimental setup and parameter settings the reader is referred to the Supplementary Material.

To evaluate the compared methods following accepted standards, the resulting gene predictions reported by the different methods were analyzed using the framework provided by Cufflinks in the analysis tool Cuffcompare (Trapnell *et al*., 2012) with the annotated coding sequences of NCBI as a reference transcript set. Here, the specificity and sensitivity for base level, exon, transcript, locus and intron level are reported, following the guidelines presented in Burset and Guigó (1996). Following this framework, we also report fuzzy measures of these quantities, which report whether correct identification were found in proximity even though the precise location might have been missed. These numbers complement the exact numbers, for instance to give an impression of how many exons have been predicted almost completely but without the exact boundary. To ensure a fair comparison between methods, we masked all direction information in the Cuffcompare analysis, as Cufflinks does not report any frame information in case no splicing events occur. In addition, we generated receiver operating characteristic (ROC) curves complemented by calculating the F-measure (van Rijsbergen, 1979) for our measures of sensitivity and specificity. For further details on the comparison framework and the calculation of sensitivity and specificity refer to the Supplementary Material.

Finally, for the two real datasets and the human simulation, we performed an alternative evaluation study based on sampling a fixed number of predictions for all compared methods. This way the measure of accuracy is independent of the overall number of predictions of each tool. The results of this evaluation are included in the Supplementary Material.

4 RESULTS

In our study, we intend to demonstrate the applicability of GIIRA on different organism types and the effect of including ambiguous mappings in the analysis. Thus, a crucial point is the proportion of ambiguously mapped reads in the alignment. All mappings showed ambiguity, although in varying levels: with 6.6% the simulated *E.coli* data had the lowest proportion of ambiguous mappings, whereas the real *E.coli* experiment showed the highest proportion with 97%. The human simulation showed 22.8% and the yeast datasets 19% ambiguous hits, respectively. Note that the large proportion on the real *E.coli* data is due to a high level of ribosomal RNA contamination within the sample (data not shown). Without contamination, the ambiguity is ~5%, similar to the mapping proportion of the simulated *E.coli* dataset that did not include contaminants. Details for all mapping results and system requirements of GIIRA are included in the Supplementary Material.

4.1 Prokaryotic datasets

Table 1 shows the Cuffcompare comparison between Cufflinks, GIIRA, GeneMark and GLIMMER3 for the *E.coli* simulation and the real *E.coli* dataset, respectively.

Because the reads were derived directly from the annotated genes, the simulation reflects the ability of the compared methods to identify expressed regions and to resolve overlaps. Cufflinks, GLIMMER3 and GIIRA yield a high accuracy on the base level, with GIIRA being more specific than Cufflinks and GLIMMER3, whereas GLIMMER3 is slightly more sensitive than GIIRA. Compared with GeneMark, all methods show a sensitivity and specificity increased by more than 20%. Because only GIIRA and GLIMMER3 focus on extracting structural genes rather than operons or expressed areas, it is not surprising that on exon and locus level both methods show clearly better accuracy than the competing methods (refer to Table S4 and Fig. S5 in the Supplementary Material). In the *E.coli* simulation, GIIRA yields more accurate results than GLIMMER3 with sensitivity and specificity increased by up to 6 and 11%, respectively, on the locus level. It should be noted that for Cufflinks only the fuzzy exon and locus level are of relevance, as Cufflinks does not predict start and stop codons and thus regularly misses bases at the start and end of genes. The fuzzy category covers these bases because here not only a perfect match, but also a match in range around the correct result is accepted. For GIIRA, the fuzzy *Sn* and *Sp* are only slightly increased compared with the perfect match *Sn* and *Sp*, indicating a high accuracy in predicting the correct frame for an expressed region.

Table 1. Cuffcompare analysis for the simulated (1) and real (2) *E.coli* dataset

Method	(1) *E.coli* simulation		(2) *E.coli* real	
	Sensitivity	Specificity	Sensitivity	Specificity
GIIRA	96.5	**97.7**	61.4	93.3
Cufflinks	91.1	92.5	40.7	72.2
GeneMark	69.2	66.5	56.1	47.9
GLIMMER3	**96.7**	94.6	**96.7**	**94.6**

Note: The highlighted numbers indicate the best results on the base level for sensitivity and specificity, respectively, for GIIRA, Cufflinks, GeneMark and GLIMMER3. Note that for the real *E.coli* dataset sensitivity and specificity values are only relative measurements to compare the four methods, but cannot be regarded as absolute numbers. As opposed to our simulation, where all annotated genes are represented, not all of the genes in *E.coli* are necessarily expressed at the same time. Thus, in particular the values for GLIMMER3, the only exclusively *ab initio* method, do not reflect the genes actually expressed but arise from the prediction of the complete set of genes.

It should be noted that for the real *E.coli* dataset *Sn* and *Sp* values are only relative measurements to compare the four methods, but cannot be regarded as absolute numbers because not all of the genes in *E.coli* are necessarily expressed at the same time. Thus, an additional analysis based on a subset of likely expressed reference genes is included in the Supplementary Material. For this dataset, it was not only important to identify expressed regions and distinguish contaminants but to also correctly predict genes within the expressed areas. This is reflected in the low sensitivity and specificity values for Cufflinks and GeneMark, as both methods have a scope differing from identifying structural genes.

As shown in Table 1 and also in Table S4 and Figure S6 in the Supplementary Material, GLIMMER3 appears to yield the best prediction accuracy on all compared levels. However, the values obtained for GLIMMER3 do not reflect the prediction of the actually expressed genes because it is the only compared method that exclusively predicts *ab initio* without including RNA-Seq evidence. For the purpose of completeness, we included the measures of sensitivity and specificity in the Cuffcompare analysis; however, in our comparison we focus on the three methods capable of RNA-Seq integration.

GeneMark and GIIRA yield comparable results on the base sensitivity level. However, GIIRA is more specific, as GeneMark covers large parts of the *E.coli* genome with operons without indicating the correct locus of the included genes. As illustrated in Figure 3 and in Table S4 and Figure S6 in the Supplementary Material, GIIRA performs better than Cufflinks and GeneMark on exon and locus level. GIIRA achieves a good prediction accuracy of the reference genes, whereas Cufflinks only predicts the expressed regions without indicating the included genes. GeneMark predicts operons, although these predicted regions also cover not expressed areas and can also span more than one operon (indicated by reference genes in different directions).

4.2 Eukaryotic datasets

Although GIIRA was primarily designed as a prokaryotic gene predictor it is also applicable to eukaryotes as examined on a simulated human and a simulated and real yeast dataset. Eukaryotic data pose challenges different from prokaryotic data; instead of distinguishing operons and determining gene structures, here many genes have alternative splice sites and

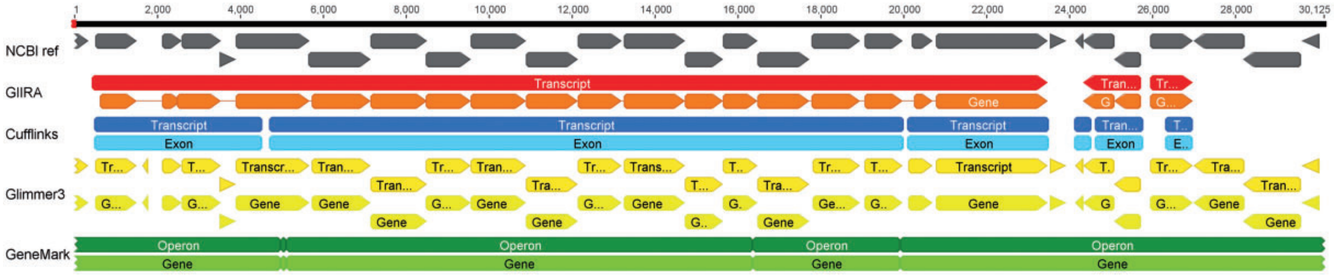

Fig. 3. Exemplary excerpt of the gene predictions of GIIRA, Cufflinks, GLIMMER3 and GeneMark for the gene region starting at position 87 000, illustrated in Geneious (Kearse *et al.*, 2012). GIIRA (transcripts in red and genes in orange) achieves a good prediction accuracy of the gray reference genes (which overlap when shown in different rows), whereas Cufflinks (blue) only predicts expressed regions without distinguishing genes. GLIMMER3 (yellow) achieves a good prediction accuracy for actually expressed genes, although it also predicts not expressed genes (e.g. on the right) because it does not consider RNA-Seq information. GeneMark (green) predicts operons, although these predicted regions also cover non-expressed areas

Table 2. Cuffcompare analysis for the simulated human data

Method	Base	Exon	Intron	Intron-chain	Transcript	Locus
			Sensitivity			
GIIRA	**97.2**	85.7	91	44.6	38.5	59.1
GIIRA_w/o	93.5	80.1	85.6	43.6	37.9	57.2
Cufflinks	93	71.6	86.7	**48.8**	0.6	56.2
AUGUSTUS	93.4	**88.6**	**91.9**	45.4	**39.3**	**59.7**
			Specificity			
GIIRA	98	**89.1**	96.7	43.3	34.9	43.9
GIIRA_w/o	**98.4**	88.4	**98.4**	44.3	34.1	39.8
Cufflinks	97.8	78.2	97.3	**51.7**	0.5	44
AUGUSTUS	82.3	81.4	85.3	49.1	**38.1**	**44.8**
			Fuzzy sensitivity			
GIIRA		**89.8**	91.7	58	**44.9**	63.5
GIIRA_w/o		84.2	86.1	53.7	43	60
Cufflinks		85.2	87.2	63.2	36	60.3
AUGUSTUS		89.4	**92.3**	**70.2**	40.6	**74.3**
			Fuzzy specificity			
GIIRA		**93.4**	97.4	56.3	**40.6**	47.1
GIIRA_w/o		92.9	**99**	54.5	38.7	41.7
Cufflinks		93	97.8	67	35.5	47.2
AUGUSTUS		82.1	85.7	**75.9**	39.4	**54.9**

Note: The highlighted numbers indicate the best results for each criterion for sensitivity and specificity for GIIRA with ambiguous reads, GIIRA without ambiguous reads (GIIRA_w/o), Cufflinks and AUGUSTUS. Note that in case of fuzzy sensitivity and specificity not only a perfect match but also a match in a range around the ground truth result is accepted.

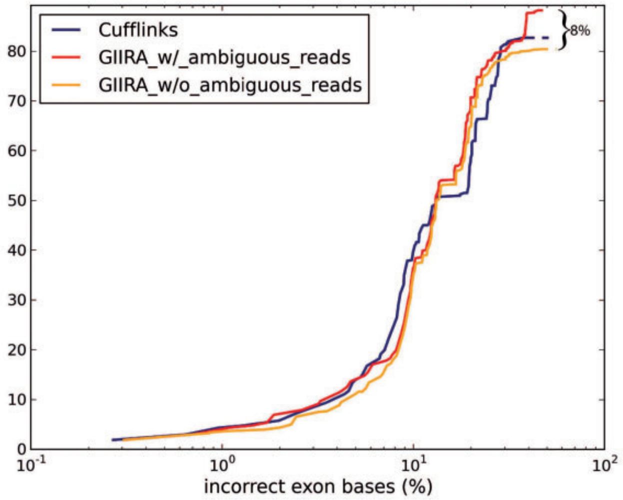

Fig. 4. ROC comparing the proportion of correctly and incorrectly predicted exonic bases for Cufflinks and GIIRA for yeast chromosome 1, with GIIRA applied in two modes: including ('GIIRA_w/_ambiguous_reads') and excluding ambiguous reads ('GIIRA_w/o_ambiguous_reads'). Including ambiguous reads increases the sensitivity by up to 8% at constant specificity. The corresponding F-measures are 75.8 for Cufflinks, 78.9 for GIIRA with ambiguous reads and 75.3 for GIIRA without ambiguous reads. Dashed lines indicate the number of bases missed due to not identifying a reference exon. Note that the proportion of false predictions is reported on a logarithmic scale

various alternative isoforms are present. In this experiment, we compared GIIRA with Cufflinks as well as AUGUSTUS as an example of a hybrid gene prediction approach. AUGUSTUS can incorporate information from RNA-Seq experiments, and here the filtered TopHat2 mapping was included according to the instructions by the authors (refer to the Supplementary Material for details). As shown in Table 2 and in Figure S7 in the Supplementary Material, GIIRA yields the most accurate predictions on the base level as well as on the fuzzy exon and transcript level, whereas Cufflinks is more accurate in predicting introns, especially exact intron-chains. Further, on the exact exon and intron level GIIRA yields a sensitivity comparable with the best values (obtained by AUGUSTUS), whereas it is clearly more specific with an increase of more than 7 and 11%, respectively. This is also reflected on the base level: here, AUGUSTUS also yields the lowest prediction specificity due to a high number of incorrectly predicted exons and their corresponding introns. However, on the locus level the hybrid prediction method AUGUSTUS outperforms GIIRA and Cufflinks by ~10% in sensitivity and 7% in specificity.

In summary, when comparing the methods that are exclusively based on RNA-Seq information, we see comparable results, with GIIRA obtaining a better sensitivity and specificity on all levels other than intron and intron-chain.

The direct comparison between GIIRA with and without ambiguous reads shows that the prediction sensitivity is increased for all levels when ambiguous mappings are included. The effect is especially pronounced on the exon and intron level, where including ambiguous reads reduces the lack of sensitivity by up to one third. Interestingly, the intron predictions become more specific when ambiguous mappings are excluded, indicating that a number of erroneous introns is due to ambiguous split reads.

As illustrated in Figure 4 and in Table S7 in the Supplementary Material, also for the *S.cerevisiae* dataset a loss in identifications can be observed when ambiguously mapped reads are disregarded, in particular the sensitivity in correctly predicted exonic bases is reduced by 8%. Overall, with more than 80% correctly predicted exonic bases GIIRA yields the highest sensitivity, whereas both Cufflinks and GIIRA are comparable in specificity. Cufflinks is more conservative than GIIRA and yields a higher number of exonic bases missed due to not predicting a complete reference exon.

The results for the simulated yeast dataset are included in the Supplementary Material. Here, the Cuffcompare analysis and ROC curve show a superior prediction accuracy of GIIRA in comparison with Cufflinks and to GIIRA excluding ambiguous reads. Only the sensitivity of intron and intron-chain predictions is slightly smaller for GIIRA than for Cufflinks.

5 DISCUSSION

We introduced GIIRA as a gene finder that identifies potential coding regions exclusively based on mapping of reads from an RNA-Seq experiment. Unlike other gene predictors, GIIRA also includes ambiguously mapped reads in the analysis, which improves on the prediction accuracy as demonstrated for various datasets with different levels of ambiguity. As shown in Section 4, already a comparably small number of ambiguous reads can substantially contribute to the ambiguity of a mapping. Disregarding this information leads to a loss in sensitivity, e.g. for genes sharing homologous regions or present in high copy numbers (refer to Section 4.2, where including ambiguous reads enhanced the sensitivity of exon predictions by up to one-third).

GIIRA accurately predicts the correct genes for prokaryotic transcripts as demonstrated in Section 4.1. It identifies the most likely set of genes explaining the expressed region using an alignment scoring adaptation coupled with a linear program formulation. In comparison with existing approaches capable of RNA-Seq integration, GIIRA has two major benefits: (i) it shows overall increased prediction accuracy and (ii) it predicts structural genes themselves rather than focusing on operons such as GeneMark or transcripts without indicating start and stop codons such as Cufflinks.

Although GIIRA was primarily designed for prokaryotic gene prediction, it can also be applied to eukaryotic gene prediction as an addition to existing annotation pipelines or a complement to other gene finders. For eukaryotic genomes, the complexity of alternative splicing events poses a critical challenge because GIIRA does not work with splice graphs to combine exons, but evaluates each splice site independently from others. As illustrated in Section 4.2, compared with the other methods GIIRA is sensitive in predicting exons and transcripts. It also yields a high accuracy in predicting introns, but is less accurate in combining them to the correct intron-chain. For instance, a challenge arises for GIIRA if two alternative isoforms share an exon where one isoform ends with this exon and the other isoform proceeds with other exons. For GIIRA, both isoforms appear to be continued with other exons and it assigns an incorrect intron-chain. Because Cufflinks uses a graph theory approach to evaluate splice sites, it is less affected by this phenomenon and on the intron-chain level it, hence, yields higher prediction accuracy than GIIRA. AUGUSTUS, as a hybrid gene predictor using non-ambiguous RNA-Seq mappings as external evidence, is less specific than the compared methods in regard to exon prediction but is superior in locus prediction.

Because GIIRA is exclusively based on RNA-Seq information, it can only predict genes currently expressed in the organism of interest and thus does not necessarily provide a complete annotation of all encoded genes.

GIIRA provides two frameworks to control the number of false-positive predictions: (i) to filter contaminants and sequencing artifacts and (ii) to verify the reported gene predictions. It can identify regions with an extremely large coverage compared with the average coverage to be sequencing artifacts or other errors such as contaminants. In case of the real *E.coli* dataset, this outlier identification filtered out most of the ribosomal RNA contaminants. Further, GIIRA reports additional information on coverage and ambiguous read support for each prediction. This enables an easy post-processing of the output allowing a trade-off of sensitivity and specificity adjusted to the intended follow-up analysis. Note that although GIIRA is independent from any *a priori* information, it is possible to use such information (if present) to improve the prediction accuracy. For instance, if a reference annotation is already available, different runs of GIIRA can be compared using the Cuffcompare framework to identify an optimal parameter setting.

6 CONCLUSION

GIIRA is a gene prediction method that identifies potential coding regions exclusively based on the mapping of reads from an RNA-Seq experiment. It was foremost designed for prokaryotic gene prediction and can resolve genes within the expressed region of an operon. However, it is also applicable to eukaryotes and predicts exon intron structures as well as alternative isoforms. Unlike other gene finders, GIIRA also incorporates ambiguously mapped reads in the gene identification, which improves the sensitivity of predictions in particular for genes sharing homologous regions or present in more than one copy on the genome. As shown for several datasets, GIIRA performs favorably in comparison with existing approaches, in particular for prokaryotes. Further, GIIRA allows an easy post-processing of the predicted genes to choose the best trade-off between sensitivity and specificity adjusted to the intended follow-up analysis.

ACKNOWLEDGEMENTS

The authors thank all members of the NG4 bioinformatics group (Robert Koch-Institute) and Wojtek Dabrowski (ZBS1, Robert Koch-Institute) for helpful and inspirational discussions and Jochen Singer (FU Berlin) for critical reading of the manuscript.

Funding: Generous start-up support by the Robert Koch-Institute (to B.Y.R.).

Conflict of Interest: none declared.

REFERENCES

Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.

Besemer,J. *et al.* (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.

Bonfert,T. *et al.* (2012) A context-based approach to identify the most likely mapping for RNA-seq experiments. *BMC Bioinformatics*, **13** (**Suppl. 6**), S9.

Burset,M. and Guigó,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.

Chung,D. *et al.* (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.*, **7**, e1002111.

CPLEX. (2011) International Business Machines Corporation. v12.4: Users manual for CPLEX. *IBM ILOG CPLEX*, IBM, Armonk, New York, United States.

Delcher,A.L. *et al.* (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.

Denoeud,F. *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, **9**, R175.

Ederveen,T.H.A. *et al.* (2013) Reduce manual curation by combining gene predictions from multiple annotation engines, a case study of start codon prediction. *PLoS One*, **8**, e63523.

Garber,M. *et al.* (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods*, **8**, 469–477.

GLPK. (2006) GNU Linear Programming Kit, v4.47. *GLPK.*

Goodswen,S.J. *et al.* (2012) Evaluating high-throughput *ab initio* gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. *PLoS One*, **7**, e50609.

Guigó,R. *et al.* (2006) EGASP: the human ENCODE genome annotation assessment project. *Genome Biol.*, **7** (**Suppl. 1**), S2.

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

Holtgrewe,M. (2010) Mason - a read simulator for second generation sequencing data. In: *Technical report TR-B-10-06*. Fachbereich für Mathematik und Informatik, Freie Universität Berlin.

Kearse,M. *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Majoros,W.H. *et al.* (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Martin,J. *et al.* (2010) Bacillus anthracis genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics*, **11** (**Suppl. 3**), S10.

Mezlini,A.M. *et al.* (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.*, **23**, 519–529.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Palmieri,N. *et al.* (2012) Evaluation of different reference based annotation strategies using RNA-Seq - a case study in *Drososphila pseudoobscura*. *PLoS One*, **7**, e46415.

Pickrell,J.K. *et al.* (2012) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Savidor,A. *et al.* (2006) Expressed peptide tags: an additional layer of data for genome annotation. *J. Proteome Res.*, **5**, 3048–3058.

Schrimpe-Rutledge,A.C. *et al.* (2012) Comparative omics-driven genome annotation refinement: application across *Yersiniae*. *PLoS One*, **7**, e33903.

Stanke,M. *et al.* (2006) Gene prediction in eukaryotes with a generalized Hidden Markov Model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stanke,M. *et al.* (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

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

Tu,Q. *et al.* (2012) Gene structure in the sea urchin *Strongylocentrotus purpuratus* based on transcriptome analysis. *Genome Res.*, **22**, 2079–2087.

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice: review of concepts, case studies and implications. *J. Mol. Biol.*, **235**, 1–12.

Wang,Q. *et al.* (2012) Theoretical prediction and experimental verification of protein-coding genes in plant pathogen genome Agrobacterium tumefaciens strain C58. *PLoS One*, **7**, e43176.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wei,C. and Brent,M. (2006) Using ESTs to improve the accuracy of *de novo* gene prediction. *BMC Bioinformatics*, **7**, 327.

Wijaya,E. *et al.* (2013) Finding protein-coding genes through human polymorphisms. *PLoS One*, **8**, e54210.

van Rijsbergen,C.J. (1979) *Information Retrieval*. 2nd edn. Butterworths, London.