# RNA-Seq Driven Gene Identification

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

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

zickmannF@rki.de
renardB@rki.de

**Abstract:** The reliable identification of genes is a challenging and crucial part of genome research. Various methods aiming at accurate predictions have evolved that predict genes *ab initio* on reference sequences or *evidence based* with help of additional information. With high-throughput RNA-Seq data reflecting currently expressed genes, a particularly meaningful source of information has become commonly available. However, a particular challenge in including RNA-Seq data is the difficult handling of ambiguously mapped reads. Therefore we developed GIIRA, a novel gene finder that is exclusively based on RNA-Seq data and inherently includes ambiguously mapped reads. 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. Further, we developed a framework to integrate GIIRA and other gene finders to obtain a verified and accurate set of gene predictions.

## 1 Introduction

Accurate gene identification is an important and also challenging part of genome analysis pipelines. Hence, various gene finders have evolved, which are categorized as *ab initio* and *evidence based* (including *comparative*) gene finders. Ab initio approaches predict genes exclusively on the target sequence and perform identifications based on training data and strategies such as Hidden Markov models. In contrast, evidence based methods report genes depending on observed evidence, such as EST libraries or protein sequences. Further, there exist hybrid approaches that combine ab initio gene prediction with additional evidence to verify the predicted genes [GKE12].

Since none of these strategies is bias-free, also methods combining predictions have evolved. These approaches introduce weighting schemes to score different predictions and merge the output of different input gene finders.
Despite all efforts, gene identification still faces significant challenges handling complex gene structures, rare splice sites or mutations in genes [GKE12]. These problems can be overcome by using the knowledge available from high-throughput RNA-Seq experiments, which reflect genes expressed under experimental conditions and hence provide valuable information to identify novel or confirm predicted genes. One challenge in including RNA-Seq information is the presence of ambiguous reads, which map to several locations in the genome and therefore complicate the gene prediction. However, discarding ambiguously mapped reads may result in a significant loss of prediction accuracy, since for instance repetitive or highly similar regions or paralogous genes lead to a substantial part of non-unique mappings.

To utilize the complete information contained in RNA-Seq experiments for gene identification we developed GIIRA (**G**ene **I**dentification **I**ncorporating **R**NA-Seq and **A**mbiguous reads), a de novo gene predictor that works on a reference genome and reads derived in a RNA-Seq experiment [ZLR14]. Since GIIRA is also excellently suited to be combined with predictions from other gene finders we further developed an easy-to-use framework to merge results of different prediction strategies. This allows to combine advantages of diverse methods to obtain a verified and accurate set of gene predictions [ZR14].

## 2 Methods and Results

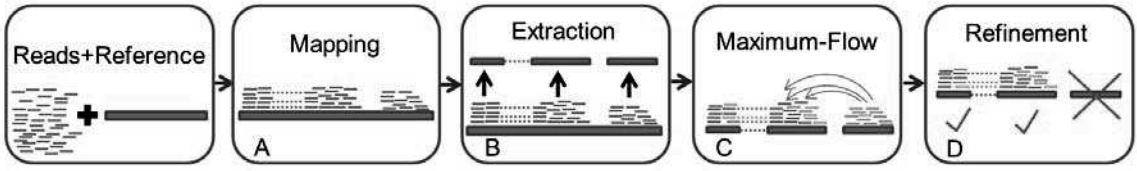


Figure 1: Workflow of GIIRA: Given a genomic sequence and a set of RNA-Seq reads, reads are mapped to the reference (A) and the resulting alignment is then analyzed by GIIRA. Candidate genes are extracted (B) and ambiguous reads are reassigned using a maximum-flow optimization (C). Finally, candidate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

As depicted in Figure 1, GIIRA is based on an alignment of reads from a RNA-Seq experiment to the DNA sequence of interest. Based on the observed mapping coverage GIIRA identifies *candidate genes* and searches the most likely start and stop position for each candidate (Fig. 1.B). Several parameters that can be user-defined or estimated from

the alignment control the verification of splicing events or coverage variations. All overlapping reads, regardless if unique or ambiguous mapping, are assigned to their corresponding candidate gene.
Prokaryotic candidates undergo an additional extraction step because prokaryotic operons contain a continuously expressed region including one or more genes that have to be identified respecting the present open reading frames (ORFs). An algorithm formulated as a linear program optimizes the set of chosen ORFs with focus on selecting a set that covers a large number of bases in this operon while restricting the overall number of ORFs.

In the 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 sequenced read can only arise from one genomic locus, we have to reassign ambiguously mapped reads to their most likely origin. To do so, GIIRA uses a maximum-flow approach formulated as an integer linear program that is based on the coverage of the gene candidates, their support from unique mappings and the ambiguity of the reads themselves (Fig. 1.C).
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 indicates 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 affects the choice for the best origin.

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 $\varphi_{ij}$ (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.
After a unique position for each read has been assigned, the candidate genes are refined accordingly and genes lacking read support are erased (Fig. 1.D). Further, all remaining genes are scored based on their read coverage and the quality and former ambiguity of assigned reads. This allows an easy post-processing to verify genes for follow-up analyses, for instance to filter out genes with overall low support.

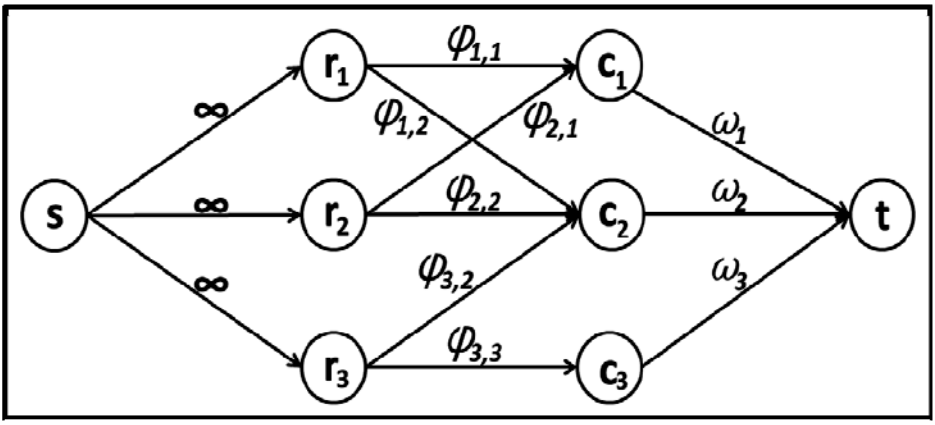


Figure 2: Simplified example for a maximum-flow network representation passing flow from source node *s* to target node *t*. The source node is connected to the nodes representing reads ($r_i$), which are connected with all genes they were mapped to ($c_j$). The edge labels indicate the capacity for 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).

We evaluated GIIRA in three simulations and two real datasets, where it performed favorably in comparison to three prokaryotic and eukaryotic gene finders as well as the RNA-Seq based method Cufflinks [TWP+10]. In particular for prokaryotes, GIIRA showed a substantial increase in both sensitivity and specificity compared to existing gene finders.
While comparing the methods, we also tested combinations of GIIRA with predictions from other gene finders. Combining the results of different gene finding strategies complemented the single method predictions and resulted in significantly improved prediction accuracy, also in comparison to existing approaches for prediction combination [ZR14].

## References

[GKE12] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis. Evaluating High-Throughput Ab Initio Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. *PLoS ONE*, 7(11), 2012.

[TWP+10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511-515, 2010.

[ZLR14] Franziska Zickmann, Martin S. Lindner, and Bernhard Y. Renard. GIIRA – RNA-Seq driven gene finding incorporating ambiguous reads. *Bioinformatics*, 30(5):606-613, 2014.

[ZR14] Franziska Zickmann and Bernhard Y. Renard. Submitted manuscript under review. 2014.