Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Data and text mining

TIminer: NGS data mining pipeline for cancer immunology and immunotherapy

Elias Tappeiner [1], Francesca Finotello [1], Pornpimol Charoentong [1], Clemens Mayer [1], Dietmar Rieder [1], and Zlatko Trajanoski [1,*]

[1] Biocenter, Division of Bioinformatics, Medical University of Innsbruck, Innsbruck, 6020, Austria.

*To whom correspondence should be addressed.

Associate Editor: Dr. Jonathan Wren

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Recently, a number of powerful computational tools for dissecting tumor-immune cell interactions from next-generation sequencing (NGS) data have been developed. However, the assembly of analytical pipelines and execution of multi-step workflows are laborious and involve a large number of intermediate steps with many dependencies and parameter settings. Here we present TIminer, an easy-to-use computational pipeline for mining tumor-immune cell interactions from NGS data. TIminer enables integrative immunogenomic analyses, including: human leukocyte antigens typing, neoantigen prediction, characterization of immune infiltrates, and quantification of tumor immunogenicity.
Availability: TIminer is freely available at http://icbi.i-med.ac.at/software/timiner/timiner.shtml
Contact: zlatko.trajanoski@i-med.ac.at.

1 Introduction

Recent breakthroughs in cancer immunotherapy and decreasing costs of next-generation sequencing (NGS) technologies sparked intensive research into tumour-immune cell interactions using genomic tools. The wealth of the generated data and the added complexity pose considerable challenges and require computational tools to process, analyze, and visualize the data. Recently, several tools and analytical pipelines have been developed and used to effectively mine tumour immunologic and genomic data and extract information relevant for cancer immunology and immunotherapy that includes: human leukocyte antigens (HLA) typing, prediction of neoantigens (non-self tumor antigens predicted from somatic mutations binding specific HLA types), tumor-infiltrating immune cells estimated from RNA sequencing (RNA-seq) data, and expression levels of key molecules like immune checkpoints (e.g. cytotoxic T-lymphocyte-associated antigen (CTLA4) and programmed cell death protein 1 (PD1)) and major histocompatibility complex (MHC) molecules (see our recent review (Hackl *et al.*, 2016)). However, to the best of our knowledge, there is currently no easy-to-use analytical pipeline to perform integrative immunogenomic analyses. Several tools have been recently published but provide only limited functionality like the identification of tumor neoantigens (Supplementary Table S1). Moreover, assembly of such analytical pipelines and execution of multi-step workflows are laborious and involves a large number of intermediate steps with many dependencies and parameter settings.

We therefore developed TIminer (Tumor Immunology miner), an analytical pipeline to perform integrative immunogenomic analyses using NGS data that is easy to install and use. TIminer integrates state-of-the-art bioinformatics tools to analyze single-sample RNA-seq data and somatic DNA mutations to characterize the tumor-immune interface including: 1) genotyping of HLAs from NGS data, 2) prediction of tumor neoantigens using mutation data and HLA types, 3) characterization of tumor-infiltrating immune cells from bulk RNA-seq data, and 4) quantification of tumor immunogenicity from expression data.

2 Tool description

TIminer considers RNA-seq reads and somatic DNA mutations to perform immunogenomic analyses (Figure 1). RNA-seq reads must be provided as FASTQ files, whereas files of somatic DNA mutations should follow the Variant Call Format (VCF). The different analyses, together with their input and output data, are described in the following.

- RNA-seq FASTQ files are used to quantify the gene expression with Kallisto (Bray *et al.*, 2016) as: gene-specific counts, transcripts per millions (TPM), and normalized $\log_2$(TPM+1).

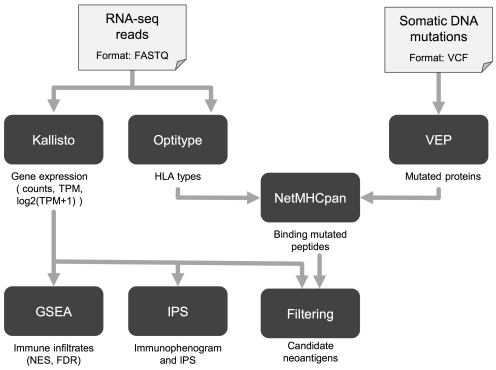

Fig. 1. The scheme illustrates the different computational tools integrated in TIminer, the input/output data, and the data flow between the tools.

- FASTQ files are analyzed with Optitype (Szolek *et al.*, 2014) to predict class-I HLA types. The output reports the 4-digit predictions of the HLA-A, HLA-B, and HLA-C alleles.
- Normalized $\log_2$(TPM+1) are used to estimate the enrichment of tumor-infiltrating immune cell types through gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) selecting between three different gene-set schemes (Angelova *et al.*, 2015; Charoentong *et al.*, 2017; Aran *et al.*, 2017) or providing a file with custom gene sets. The output files report the normalized enrichment scores (NES) and the associated false-discovery-rate (FDR)-corrected p-values.
- Normalized $\log_2$(TPM+1) are also used to depict the major determinants of tumor immunogenicity through an immonophenogram and to compute the immunophenoscore (IPS), representing the overall tumor immunogenicity (Charoentong *et al.*, 2017).
- Single-nucleotide DNA mutations affecting coding regions are annotated by VEP (McLaren *et al.*, 2016) and used to predict the sequences of the affected proteins.
- The mutated proteins arising from missense mutations, together with the predicted class-I HLA types, are subjected to NetMHCpan (Nielsen and Andreatta, 2016) to predict mutated peptides binding to HLA molecules.
- Binding peptides are filtered considering TPM to select only candidate neoantigens arising from expressed genes. Alternatively, sensitive filtering of neoantigens can be performed considering allele-specific gene expression (see Supplementary Note).

3 Implementation

The computational tools performing the immunogenomic analyses integrated in the pipeline have different dependencies, interfaces, and input/output data formats. The installation and assembly of such tools is very laborious and time consuming. TIminer was explicitly designed and implemented to overcome these drawbacks.

The software was pre-built into a ready-to-use Docker image to speed up and simplify the installation process, thereby preventing possible dependency issues and version conflicts. The TIminer framework can be installed on Unix operating systems, including Mac OS, using a one-click installer. In order to simplify the access to the individual tools, we provide a unified Python application programming interface (API). The TIminerAPI wraps the tool calls into the Docker image, introduces parallel execution traces to optimize the computation, and handles all required data pre- and post-processing steps. The single tools are accessible over the TIminerAPI individually or can be run together as part of the full pipeline, available as a python script. In addition, TIminer encloses a graphical user interface (GUI) that enables single-patient data analysis on standard desktop computer.

Given the size of NGS input data, we recommend analysis of large cohorts of patients with TIminer on computing cluster units, which are available in many institutions. Single-patient data can be analyzed on desktop computers or laptops. Example files, provided within the TIminer package, can be used to test the full pipeline (see Supplementary Note).

4 Conclusions

We present here TIminer, the first analytical pipeline that performs integrative immunogenomic analyses from NGS data including HLA typing, neoantigen prediction, characterization of immune infiltrates, and quantification of tumor immunogenicity. Moreover, the pre-built and ready-to-use Docker image enables simple installation procedure. Hence, TIminer represents a valuable tool for basic and translational research in cancer immunology and can expedite the development of precision immuno-oncology. Although developed for cancer immunology and immunotherapy, TIminer provides the means to study also autoimmune, inflammatory, or infectious diseases.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF): W1101-B18 (DK Molecular Cell Biology and Oncology), the European Union's Horizon 2020 research and innovation program under grant agreement No. 633592 (project APERIM: Advanced Bioinformatics Tools for Personalised Cancer Immunotherapy), and the Austrian National Bank (Jubiläumsfondsprojekt No. 16534).

References

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., Waldner, M. J., Bindea, G., Mlecnik, B., Galon, J., and Trajanoski, Z. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biology*, **16**, 64.

Aran, D., Hu, Z., and Butte, A. J. (2017). xcell: Digitally portraying the tissue cellular heterogeneity landscape. *bioRxiv*, page 114165.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, **34**(5), 525–527.

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H., and Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Reports*, **18**(1), 248–262.

Hackl, H., Charoentong, P., Finotello, F., and Trajanoski, Z. (2016). Computational genomics tools for dissecting tumour-immune cell interactions. *Nature Reviews. Genetics*, **17**(8), 441–458.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, **17**(1), 122.

Nielsen, M. and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Medicine*, **8**(1), 33.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550.

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*, **30**(23), 3310–3316.