OXFORD

Genome analysis

FUNGI: FUsioN Gene Integration toolset

Alejandra Cervera [1,2,†], Heidi Rausio[3,†], Tiia Kähkönen[3], Noora Andersson[4], Gabriele Partel[1], Ville Rantanen[1], Giulia Paciello[5], Elisa Ficarra [6], Johanna Hynninen[7], Sakari Hietanen[7], Olli Carpén[1,3,4], Rainer Lehtonen[1], Sampsa Hautaniemi[1,*] and Kaisa Huhtinen[1,3,*]

[1]Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, 00014 Helsinki, Finland, [2]Laboratorio de Diabetes, Facultad de Química, Universidad Nacional Autónoma de México, Campus Yucatán, 97302 Mérida, México [3]Cancer Research Unit, Institute of Biomedicine and FICAN West Cancer Centre, University of Turku, 20014 Turku, Finland, [4]Department of Pathology, University of Helsinki and HUS-Diagnostics, Helsinki University Hospital, 00014 Helsinki, Finland, [5]Department of Computer Engineering, University of Modena and Reggio Emilia, 41121 Modena, Italy, [6]Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia (UNIMORE), 42121 Reggio Emilia, Italy and [7]Department of Obstetrics and Gynecology, University of Turku and Turku University Hospital, 20521 Turku, Finland

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Abstract

Motivation: Fusion genes are both useful cancer biomarkers and important drug targets. Finding relevant fusion genes is challenging due to genomic instability resulting in a high number of passenger events. To reveal and prioritize relevant gene fusion events we have developed FUsionN Gene Identification toolset (FUNGI) that uses an ensemble of fusion detection algorithms with prioritization and visualization modules.

Results: We applied FUNGI to an ovarian cancer dataset of 107 tumor samples from 36 patients. Ten out of 11 detected and prioritized fusion genes were validated. Many of detected fusion genes affect the PI3K-AKT pathway with potential role in treatment resistance.

Availability and implementation: FUNGI and its documentation are available at https://bitbucket.org/alejandra_cervera/fungi as standalone or from Anduril at https://www.anduril.org.

Contact: sampsa.hautaniemi@helsinki.fi or kaisa.huhtinen@utu.fi

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

1 Introduction

Chromosomal aberrations, such as translocations, deletions and chromosomal inversions, are common in all cancers. An important consequence of such aberrations is the fusing of two normally unrelated gene DNA segments, called fusion genes. As fusion genes are rarely present in non-malignant cells, they are considered as ideal targets for therapeutics and diagnostics purposes (Mertens et al., 2015). Fusion gene discovery has exploded due to next-generation sequencing with over 10 000 fusions currently identified. Although most of these fusions are not recurrent and are most likely passenger events, for example, fusions in ALK, ROS1, or PDGFB, or RET, NTRK1/2/3, FGFR1/2/3 and BRAF/CRAF either have targeted therapy available or are currently being tested (Schram et al., 2017).

To facilitate identification of recurrent fusion genes and fusion genes with already known therapies, we have developed FUsionN

Gene Identification toolset (FUNGI). While several RNA-seq fusion calling methods have been developed (Haas et al., 2019), no gold standard for fusion gene detection exists and it has been suggested that fusion gene calling would benefit from an ensemble of methods (Gao et al., 2018). FUNGI uses database annotation from FusionCatcher (Nicorici et al., 2014) and expands its application to the combined list of fusions and complements these database annotations by matching fusions to several Ensembl's tables (paralog, homolog and GO terms) (Cunningham et al., 2019). FUNGI also provides scoring of oncogenic potential of fusions by both Pegasus (Abate et al., 2014) and Oncofuse (Shugay et al., 2013). For visualization, we have implemented a custom-made script for recreating the exact fusion and mapping the reads to it, as well as support for TrinityFusion (FusionInspector) (Haas et al., 2019) for de novo assembly of fusions and remapping.

Similarly to FUNGI, FusionHub (Panigrahi et al., 2018) provides fusion database annotation, scoring with Oncofuse and visualization. The main difference with FUNGI is that FusionHub is an online resource for annotation that does not provide fusion calling, and that additionally it provides siRNA prediction based on the fusions. FUNGI is similar to nextflow based rnafusion (https://github.com/nf-core/rnafusion) in the capacity to run several fusion calling algorithms, integrating them and generating a report, but rnafusion lacks oncogenic scoring methods.

2 Implementation

FUNGI's steps are implemented as four separate modules, which allows flexible use of the framework: modules can be run sequentially as a pipeline or as part of workflows in which users can incorporate novel fusion calling algorithms or filtering steps in between modules. The four main modules are described below, and an overview of FUNGI is shown in Figure 1.

FusionCaller identifies gene fusion events using RNA-Seq data. It currently supports six different fusion calling algorithms: STAR-Fusion (Haas et al., 2017), FusionCatcher, ChimeraScan (Iyer et al., 2011), EricScript (Benelli et al., 2012), SoapFuse (Jia et al., 2013) and Arriba (https://github.com/suhrig/arriba/). The output folder produced by FusionCaller can be passed on to the FUNGI's FusionAnalyzer module directly or after method-specific filtering.

FusionAnalyzer filters, annotates and scores fusions coming directly from the FusionCaller module or from any other method once

transformed into a standard format (example in Supplementary Data). Next, reported breakpoints are queried in Ensembl database to corroborate that the each detected fusion matches existing genes, to annotate fusion genes with Ensembl ids and to identify among them pairs of paralogs, homologs and genes with no known function to be (optionally) discarded. Remaining fusions are checked against known fusions datasets for exclusion of common artifacts or fusions previously reported in healthy individuals. After filtering, fusions are scored by both Pegasus and Oncofuse. The output from both algorithms is combined in a final report with the most relevant information from both methods. Each filtering step keeps a log of the fusions and the reasons for exclusion at every point.

FusionVisualizer makes a virtual reference of the fusions both with or without breakpoint information. When the coordinates of the breakpoint in each gene are provided, each chimeric transcript is reconstructed from the fusion information and then reads are mapped to it. Alternatively, FUNGI uses FusionInspector (part of STAR-Fusion) for assembling new chimeric transcripts using only the gene names and aligning reads to it. The virtual references can be loaded to a genome browser, such as IGV (Thorvaldsdóttir et al., 2013), to inspect the quality of the alignments. FusionInspector is less sensitive than normal fusion calling algorithms (Haas *et al.*, 2019), so it should be taken into account that a fusion not confirmed by supervised fusion calling is not necessarily a false positive. FusionVisualizer's breakpoint-mode, recreating the exact fusion and mapping against it, provides useful information for manual inspection of the fusion and helps determine if it is an artifact or not. The main purpose of FusionVisualizer is to aid in rapid inspection of fusions in genome browsers and supervised fusion calling, but it is not intended
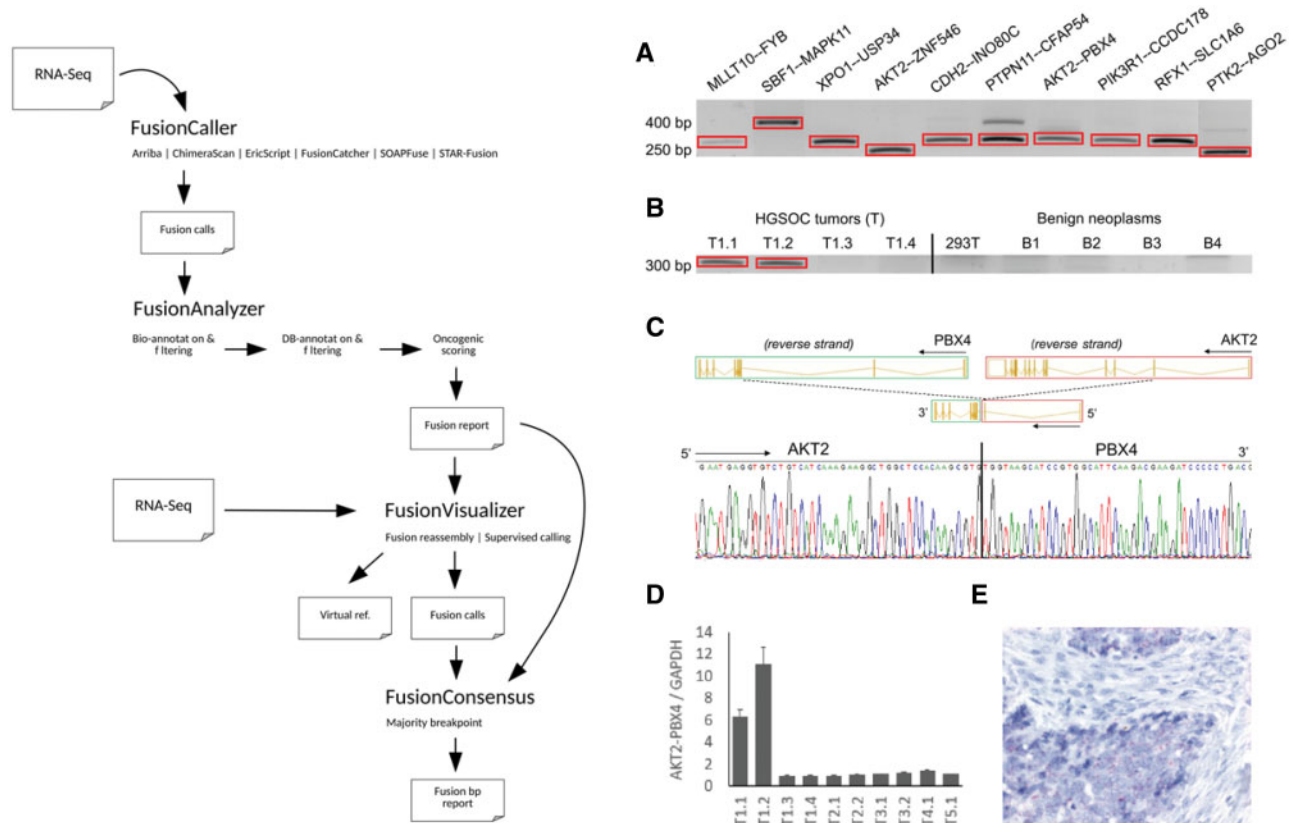
Fig. 1. FUNGI overview and experimental validation. Left panel: FUNGI's modules in a pipeline. Right panel: (**A**) Validation of ten gene fusions at mRNA level in the samples suggested by our computational fusion gene analysis pipeline. The fusion sites were amplified from HGSOC tissue cDNA using fusion-specific PCR and Sanger sequencing. (**B–E**) Example of further evaluation of AKT2-PBX4 fusion. (B) RT-PCR verified AKT2-PBX4 mRNA in the same two HGSOC tumor samples (T) as suggested by the RNA-seq data and in none of the benign (B) ovarian tumors sustaining pipeline specificity. (C) Sequence trace shows the specific nucleotide breakpoint of AKT2-PBX4 fusion in cDNA leading to fusion of exons 1–2 of AKT2 to exons 3–7 of PBX4. (D) RT-qPCR uncover expression of AKT2-PBX4 fusion in two out of four tumors (T1.1–T1.4) of the patient T1 as expected by the pipeline. Total of 10 individual tumors from five patients (T1–T5) were analyzed. (E) RNA in situ hybridization for AKT-PBX4 fusion reveals red fusion signal in the cancer cells while the surrounding stroma is free from the staining

to replace more sophisticated visualization tools such as Clinker (Schmidt *et al.*, 2018) and INTEGRATE-Vis (Zhang *et al.*, 2017).

FusionConsensus helps identify the best breakpoint for each fusion when there is no consensus between the tools. The breakpoint supported by the majority of methods is reported. FusionConsensus also combines fusion results from FusionAnalyzer and FusionVisualizer.

3 Results

We applied FUNGI to 107 RNA-seq tumor samples from 36 high-grade serous ovarian cancer (HGSOC) patients. Patient characteristics and sample processing are described in Supplementary Data. We detected 218 261 fusions, each breakpoint in each sample counted separately, that were further filtered into 228 fusion gene pairs (Supplementary Table S2). From the 228 candidate fusions, we selected 11 biologically interesting fusions for validation with cDNA Sanger sequencing (Fig. 1A, primers are listed in Supplementary Table S3). We confirmed 10 out of 11 fusions in the tumor tissues but none in benign tumors (Fig. 1B), which demonstrates that FUNGI can detect and prioritize biologically relevant fusion events when such exist in the data. RT-qPCR and RNA in situ hybridization verified the expression of AKT2-PBX4, AKT2-ZNF546 and PIK3R1-CCDC178 fusions in tumor specimens (Fig. 1D and E, Supplementary Fig. S3).

Interestingly, four of the validated fusions, i.e. AKT2-PBX4, AKT2-ZNF546, PTK2-AGO2 and PIK3R1-CCDC178, potentially affect PI3K-AKT-mTOR signaling cascade, which is constitutively hyper-activated in ∼50% of HGSOC (Mabuchi *et al.*, 2015). Furthermore, for 21 fusions a gene partner belongs to PI3K-AKT-mTOR signaling cascade or regulate it (Supplementary Table S4). Three of these fusions, FGFR3-TACC3, PTK2-AGO2 (as PTK2-EIF2C2) and FTO-RBL2, have been previously reported in the Tumor Fusion Gene Data Portal (TumorFusions, https://www.tumorfusions.org/) (Hu *et al.*, 2018), a web resource with fusions identified in The Cancer Genome Atlas (TCGA) dataset. For most of the remaining fusions ($n = 19$), the PI3K-pathway involved gene has been observed in a fusion with a different partner in TCGA. FGFR3-TACC3, was initially identified in glioblastoma and estimated to be present in 8% of patients. In gynecological cancers it has been reported in cervical cancer, but has not been found in breast cancer or in ovarian cancer (Costa *et al.*, 2016). The FGFR3-TACC3 translocation predicts sensitivity to erdafitinib in urothelial cancer (Karkera *et al.*, 2017).

We also used FusionVisualizer to hunt for 24 specific fusions from our result set on TCGA ovarian cancer dataset (TCGA Research Network, 2011) and we were able to reliably detect 11 fusions (FFPM≥0.1, JRC≥3; 6 additional fusions were identified below those thresholds) in several samples (Supplementary Table S5).

Taken together, we have shown the utility of FUNGI in detecting relevant fusion events from RNA-seq data. FUNGI is easy to install and has a modular design that facilitates the use of the supported fusion calling algorithms and oncogenic scoring methods, as well as integration of novel methods such as DeepPrior (Lovino et al., 2020) for oncogenic scoring. FUNGI's Anduril (Cervera et al., 2019) implementation includes scripts for automated installation of the third-party software used. For the standalone version we have made available a Dockerfile with working testcases from which all installation steps can be consulted.

Acknowledgement

Funding

Conflict of Interest: none declared.

Data availability

The raw sequencing data underlying this article cannot be shared publicly due to patient confidentiality restrictions. Download links to all fusion calls and intermediate results are available from https://bitbucket.org/alejandra_cervera/fungi/.

References

Abate,F. *et al.* (2014) Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst. Biol.*, **8**, 97.

Benelli,M. *et al.* (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics*, **28**, 3232–3239.

Cerami,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.

Cervera,A. *et al.* (2019) Anduril 2: upgraded large-scale data integration framework. *Bioinformatics*, **35**, 3815–3817.

Costa,R. *et al.* (2016) FGFR3-TACC3 fusion in solid tumors: mini review. *Oncotarget*, **7**, 55924–55938.

Cunningham,F. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.

Gao,Q. *et al.*; Cancer Genome Atlas Research Network. (2018) Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.*, **23**, 227–238.e3.

Haas,B.J. *et al.* (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**, 213.

Haas,B.J. *et al.* (2017) STAR-fusion: fast and accurate fusion transcript detection from RNA-Seq. bioRxiv, 120295.

Hu,X. *et al.* (2018) TumorFusions: an integrative resource for cancer-associated transcript fusions. *Nucleic Acids Research*, **46**, D1144–D1149.

Iyer,M.K. *et al.* (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, **27**, 2903–2904.

Jia,W. *et al.* (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol.*, **14**, R12.

Karkera,J. D. *et al.* (2017) Oncogenic Characterization and Pharmacologic Sensitivity of Activating Fibroblast Growth Factor Receptor (FGFR) Genetic Alterations to the Selective FGFR Inhibitor Erdafitinib. Molecular Cancer Therapeutics, **16**, 1717–1726.

Lovino,M. *et al.* (2020) DEEPrior: a deep learning tool for the prioritization of gene fusions. *Bioinformatics*, **36**, 3248–3250.

Mabuchi,S. *et al.* (2015) The PI3K/AKT/mTOR pathway as a therapeutic target in ovarian cancer. *Gynecologic Oncology*, **137**, 173–179.

Mertens,F. *et al.* (2015) The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer*, **15**, 371–381.

Nicorici,D. *et al.* (2014) FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. doi:10.1101/011650.

Panigrahi,P. *et al.* (2018) FusionHub: a unified web platform for annotation and visualization of gene fusion events in human cancer. *PLoS One*, **13**, e0196588.

Schmidt,B.M. *et al.* (2018) Clinker: visualizing fusion genes detected in RNA-seq data. *GigaScience*, **7**, giy079.

Schram,A.M. *et al.* (2017) Fusions in solid tumours: diagnostic strategies, targeted therapy, and acquired resistance. *Nat. Rev. Clin. Oncol.*, **14**, 735–748.

Shugay,M. *et al.* (2013) Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics*, **29**, 2539–2546.

The Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.

Zhang,J. *et al.* (2017) INTEGRATE-Vis: a tool for comprehensive gene fusion visualization. *Sci. Rep.*, **7**, 17808.