OXFORD (GIGA)n SCIENCE

GigaScience, 2017, 1–8

doi: xx.xxxx/xxxx
Manuscript in Preparation
Data Note

DATA NOTE

Massive NGS Data Analysis Reveals Hundreds Of Potential Novel Gene Fusions in Human Cell Lines

Silvia Gioiosa[1,4,†], Marco Bolis[2,†], Tiziano Flati[1,4], Annalisa Massini[3], Enrico Garattini[2], Giovanni Chillemi[1], Maddalena Fratelli[2,*] and Tiziana Castrignanò[1,*]

[1]SCAI-Super Computing Applications and Innovation Department, CINECA, Rome, Italy, and [2]Laboratory of Molecular Biology, IRCCS-Istituto di Ricerche Farmacologiche "Mario Negri," Milano, Italy, and [3]Computer Science Department, Sapienza University of Rome, Italy, and [4]National Council of Research, CNR, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Bari, Italy.

[†]Contributed equally.

[*] To whom correspondence should be addressed: t.castrignano@cineca.it; maddalena.fratelli@marionegri.it

Abstract

Background: Gene fusions derive from chromosomal rearrangements and the resulting chimeric transcripts are often endowed with oncogenic potential. Furthermore, they serve as diagnostic tools for the clinical classification of cancer subgroups with different prognosis and, in some cases, they can provide specific drug targets. So far, many efforts have been carried out to study gene fusion events occurring in tumor samples. In recent years, the availability of a comprehensive Next Generation Sequencing dataset for all the existing human tumor cell lines has provided the opportunity to further investigate these data in order to identify novel and still uncharacterized gene fusion events.

Results: In our work, we have extensively reanalyzed 935 paired-end RNA-seq experiments downloaded from "The Cancer Cell Line Encyclopedia" repository, aiming at addressing novel putative cell-line specific gene fusion events in human malignancies. The bioinformatics analysis has been performed by the execution of four different gene fusion detection algorithms. The results have been further prioritized by running a bayesian classifier which makes an *in silico* validation. The collection of fusion events supported by all of the predictive softwares results in a robust set of ~ 1,700 *in-silico* predicted novel candidates suitable for downstream analyses. Given the huge amount of data and information produced, computational results have been systematized in a database named LiGeA. The database can be browsed through a dynamical and interactive web portal, further integrated with validated data from other well known repositories. Taking advantage of the intuitive query forms, the users can easily access, navigate, filter and select the putative gene fusions for further validations and studies. They can also find suitable experimental models for a given fusion of interest.

Conclusions: We believe that the LiGeA resource can represent not only the first compendium of both known and putative novel gene fusion events in the catalog of all of the human malignant cell lines, but it can also become a handy starting point for wet-lab biologists who wish to investigate novel cancer biomarkers and specific drug targets.

Key words: Database; Human gene fusions; Malignant Cell Lines; NGS; Gene Fusion detection algorithms; Chromosomal rearrangements; Bioinformatics

Background

Oncogenic gene fusion events result from chromosomal rearrangements which lead to the juxtaposition of two previously separated genes. The accidental joining of DNA of two genes can generate hybrid proteins. It can also result in the misregulation of the transcription of one gene by the *cis-regulatory* elements (promoters or enhancers) of another, sometimes resulting in the production of oncoproteins that bring the cell to a neoplastic transformation 1. Not only gene fusions can have a strong oncogenic potential 2, but they also serve as diagnostic tools for the clinical classification of cancer subgroups with different prognosis and, in some cases, they may

- A massive bioinformatics analysis conducted on Paired-End RNA-seq samples from 935 human malignant Cell Lines reveals a landscape of known and novel *in-silico* predicted gene fusion events;
- LiGeA Portal represents a user-friendly database for the systematization, visualization and interrogation of the results;
- LiGeA Portal is further integrated with information from other databases and with gene-fusion priotirization analysis, in order to address targeted experimental validations on a highly reliable set of candidate gene fusions.

provide specific drug targets 3. For instance, the presence of the PLM-RARA fusion product is a specific hallmark of acute promyelocytic leukemia (APL) 4 and represents the first example of gene-fusion targeted therapy 5 that has changed the natural history of this disease. Hence, there are several reasons why studying gene fusions in cancer is very important. In recent years, Next-Generation Sequencing (NGS) technologies have played an essential role in the understanding of the altered genetic pathways involved in human cancers. Nowadays, most of the studies aiming at fusion discovery use NGS techniques followed by massive bioinformatics analyses. The greatest challenge of these sophisticated algorithms of prediction is the ability to discriminate between artifacts and really occurring chromosomal rearrangements 6. Moreover, each gene fusion predicting software differs in terms of sensitivity and specificity. In the last decade, much effort has been done to catalog gene fusion events, thus resulting in a wide production of databases. At present, a dozen of published databases regarding oncogenic fusion genes exists (see table 1 for a summary). Some of them (e.g. FusionCancer, ChiTaRS-3.1) collect *in silico* predictions of chimeric genes, obtained analyzing publicly available datasets derived from heterogeneous sources either in terms of experimental material (a mix of Single-End and Paired-End RNA-seq data, ESTs) and in terms of data source (patients and cell lines). Some others collect gene fusion events with experimental evidences manually curated from literature collection (e.g. TCGA, Mitelman, TICdb, COSMIC, ONGene). In this work we focused on the whole catalog of Human malignant Cell Lines, thus obtaining a homogeneous input NGS dataset covering several human malignancies. We exerted a massive bioinformatics analysis on 935 paired-end RNA-seq samples derived from 22 different tumor tissues and used a combination of the best performing gene fusion-detecting algorithms. For ease of understanding, we define the predicted Gene Fusion Event (pGFE) as the entity constituted by the gene fusion couple in a specific cell line and designate the Consensus Call-Set (CCS) as the number of pGFEs supported by all the used algorithms. Starting from this assumption, we obtained a total of 377,540 pGFEs, 2,521 of which belonging to the CCS. Moreover, since not all the pGFEs can give rise to oncogenic transformations, the use of a priotirization software is recommended in order to distinguish between real driver mutations from passenger ones. Therefore, a robust Bayesian classifier has been used to perform an *in silico* validation of the results. Since one of the main purposes of this big data analysis is encouraging the reuse of our results in order to experimentally validate the *in-silico* predictions, we set up a web portal collecting and systematizing these data, LiGeA (cancer cell LInes Gene fusion portAl). It is possible to browse, search and freely download all the results obtained and described within this article at the LiGeA repository web page available at http://hpc-bioinformatics.cineca.it/fusion/. To our knowledge, our resource represents the first compendium of both known and predicted novel gene fusion events in cell lines from 22 different human tumor types.

Data Description

Methods

We have analyzed 935 paired-end RNA-seq experiments available at the Cancer Cell Line Encyclopedia repository 15, for a total of 32 TB of input raw data. The analysis has been carried out by using four different somatic fusion gene detection algorithms: FusionCatcher 16, EricScript 17, Tophat-Fusion 18 and JAFFA 19. The choice of the algorithms was driven by the assessment from Kumar S. et al. 20, which compared twelve methods for the fusion transcripts detection from RNA-Seq data and identified these softwares as the ones with the highest Positive Prediction Values. Furthermore, the chosen softwares differ in a variety of aspects and contain several layers of information in their output files, thus giving us the opportunity to collect and interconnect a wide set of complementary data for each pGFE. Here is a short description of each fusion detection tool, accompanied by the used versions and parameters.

- **FusionCatcher (FC)**: FC is a Python based algorithm. It executes a first mapping run with Bowtie v.1.2.0 21 and then performs the Gene fusion detection basing on three different aligners: Bowtie2 v.2.2.9 22, BLAT v.36 23 and STAR v.2.5.2b 24. FC takes advantage of NCBI Viral Genomes (v. 2016-01-06) in order to detect exogenous virus material integration into the host genome. Moreover, the FC algorithm compares its own output with a set of published databases, thus proving a detailed list of truly positive and false positive pGFEs candidates. In our analysis we downloaded FC v. 0.99.5a and Ensembl genome annotation v.83 and used hg38/GRCh38 as genome assembly version. The software was executed with default parameters, requiring 111,620 CPU core hours, 125 GB of RAM and 20 CPUs to complete the execution on our input dataset. Overall, FC detected 25,251 pGFEs involving 8,659 genes.
- **Tophat-Fusion (TF)**: TF uses the Tophat-fusion-post function in order to create a filtered list of gene fusion candidates, starting from the output files obtained running Tophat with the "–fusion-search" option 25." The following commands were run subsequently:

```
tophat -o $Sample.output/ -p 20 –fusion-search –keep-fasta-order –bowtie1 –no-coverage-search -r 160 –mate-std-dev 34 –max-intron-length 100000 –fusion-min-dist 100000 –fusion-anchor-length 13 $BOWTIE_INDEX/hg38 $Sample_1.fastq $Sample_2.fastq

cd $Sample.output/

tophat-fusion-post -p 20 –skip-blast $BOWTIE_INDEX/hg38
```

  Tophat-2.0.12 and samtools 0.1.19 versions were used for this study. This algorithm took about 200,000 CPU core hours, 20 CPUs and 125 GB of RAM in order to complete its runs on the whole input dataset. TF produces several output files but only the file named "results.txt", representing the filtered list of predicted gene fusions, was used for subsequent analysis. The results encompassing "Chromosome M" have been manually discarded from the final results, *in primis* because TF and JF were the only ones of the four algorithms reporting them, secondly because they represented *bona-fide* false positive outcomes. Overall, TF highlighted 28,146 pGFEs involving 9,492 genes.
- **JAFFA (JA)**: JAFFA (v. 0.9) is a multi-step pipeline that

Compiled on: June 1, 2018.
Draft manuscript prepared by the author.

takes raw RNA-Seq reads and outputs a set of candidate fusion genes along with their cDNA breakpoint sequences. It relies on trimmomatic 26, samtools 27, BLAT 23, bowtie2, bpipe 28 and R softwares 29 as well as on gencode (v. 22) for the annotation and on Mitelman database for flagging already known gene fusions. For the purpose of this analysis, we used the "Direct" mode pipeline which is indicated for reads of 100 bp or longer. A total amount of 1,300,000 CPU core hours, 125 GB of RAM and 20 CPUs were required to successfully complete the analysis. The results encompassing "Chromosome M" have been manually discarded from the final results. Furthermore, only pGFEs supported by at least 3 spanning reads or flagged as "known", have been retained. Overall, after the filtering process, JA detected 53,400 pGFEs involving 12,256 genes.

- **EricScript (ES)**: ES is developed in R, perl and bash scripts. It uses the BWA aligner 30 to perform the mapping on the transcriptome reference and samtools v. 0.1.19 to handle with SAM/BAM files. Recalibration of the exon-junction reference is performed by using BLAT. For the purposes of this project, we used BLAT v.36, R v.3.3.1, bedtools v. 2.24, and ES version 0.5.5. The Ensembl Database v. 84 was obtained as ES supplementary material 31 and built locally using BWA software with the command:

  ```
  bwa index -a bwtsw allseq.fa
  ```

  A total amount of 130,900 CPU core hours, 125 GB of RAM and 20 CPUs were required to successfully complete the analysis. We further filtered out ES final results by removing all the predictions for which the software was not able to predict an exact breakpoint position because such pGFEs could not even be experimentally validated. Secondly, as also applied to FC, TF and JF's results, we retained the pGFEs exhibiting at least 3 spanning reads over the gene fusion junction. Furthermore, we filtered out all the pGFEs with EricScore value less than 0.85. EricScore is a ranking parameter ranging from 0.5 to 1: greater values correspond to better predictions. Interestingly, by applying these filters, we filtered out almost 2/3 of the initial predictions from EricScript but, at the same time, the CCS did not reduce substantially, thus indicating that the choice of a consensus of predictions is a good strategy to remove false positives and obtain a reliable set of gene fusion candidates to be experimentally validated. Overall, after the filtering process, ES detected 293,220 pGFEs involving 14,740 genes.

Data Statistics and Validation

Overall, our extensive analysis results in a CCS of 2,521 pGFEs (Fig. 1A) and respectively 2,828/9,258 pGFEs supported by exactly three/two methods. As a first validation of our analysis, 661 out of the 719 (92%) genes known to be functionally implicated in cancer and collected under COSMIC gene census, are present in our final dataset. As a further validation of our results, about 1/5 of our CCS has already been published or is present in the following databases: chimerdb3; ONGene; COSMIC; tcga; ticdb; Mitelman (Fig. 1C). Finally, only a small subset of the pGFEs (~10% of data) present in the CCS have been recognized as false positive predictions, thus supporting the idea that a combination of algorithms can be of great utility in order to increase the sensitivity and the specificity of the tests. It is worth mentioning that, not only our analysis confirmed a large number of known gene fusion events, but it also highlighted 1,719 novel putative pGFEs in the CCS which could undergo further downstream analysis (Fig. 1B). Therefore, a further step of analysis was run with Oncofuse v.1.1.1 32 in order to distinguish driver mutations (genomic abnormalities responsible for cancer) from passenger ones (inert somatic mutations not implicated in carcinogenesis). Oncofuse is considered an *in silico* validation post-processing step which prioritizes the results obtained from each of the three algorithms. It assigns a functional prediction score to each putative fusion sequence breakpoint identified by the four softwares thus hinting which pGFEs are worthy of being experimentally validated and studied. Oncofuse supports multiple input formats such as the output from TF and FC. In order to run it also on the outputs from ES and JF, a short pre-processing step was executed on these data. As suggested on Oncofuse manual, the accepted default input format is a tab-delimited file with lines containing 5' and 3' breakpoint positions. Therefore, these columns were extracted from ES and JF output files and redirected into Oncofuse accepted input format. Oncofuse was run with default parameters using hg38 as the reference genome.

Availability of supporting data and materials

The datasets obtained and described within this article are freely downloadable at the LiGeA repository available at http://hpc-bioinformatics.cineca.it/fusion/downloads. Moreover, archival copies of processed files and the source code are available via the GigaScience database, GigaDB 33.

Database Description

LiGeA is a database server based on graph-db technology (Neo4j). The portal stores all of the results obtained from each fusion gene predicting algorithm and the prioritization analysis outcome. Anyway, this database contains not only a mere collection of *in silico* predictions. Indeed, it has been integrated with other useful external resources in order to offer a carefully-curated web compendium. Here is a short list of the added features:

- Whenever the gene fusion couple has already been experimentally validated and published, an extra column with COSMIC icon is added to the results. By clicking on it, the user will be redirected to an external link containing a manually-curated catalog of 212 literature-derived somatic mutations in cancer 34;
- Cancer Gene Census is a manually curated catalog of 719 genes for which mutations have been causally implicated in oncogenesis 35. Whenever one of the two genes involved in the pGFE has been already described to be implicated in cancer, the gene is tagged with an icon. By clicking on it, an external link to the Cancer Gene Census is provided showing a table of genes included within this category 36.
- A legend based on a colorful signature has been added to tag the FC predictions as 'validated truly positive couples' (green circle), 'validated false positive couples' (red circle), 'false positive couples with medium probability' (orange circle) and 'ambiguous signature' because tagged with both positive and negative values (grey circle) ;
- A functional prediction score obtained by extensively running the Oncofuse software, is reported as additional tag to the outputs from each algorithm.

LiGeA portal is divided into several sections which allow a user-friendly navigation.

- **Home:** In the homepage, the user is provided with a quick overview of the database. A global summary table reports a numeric recapitulation (e.g. the number of genes/transcripts/exons collected into the portal; the number of predicted proteins and so on). Moreover, a histogram shows an abstract of the top 50 involved cell lines. By moving the cursor on the bars, a pop-up opens showing the cell line name and the corresponding number of the unique fusion events predicted by all the algorithms. Information about the algorithm predictions hosted into the portal are supplied with an interactive Venn Diagram linked to a dynamical table. Upon user selection of the algorithm/s of in-

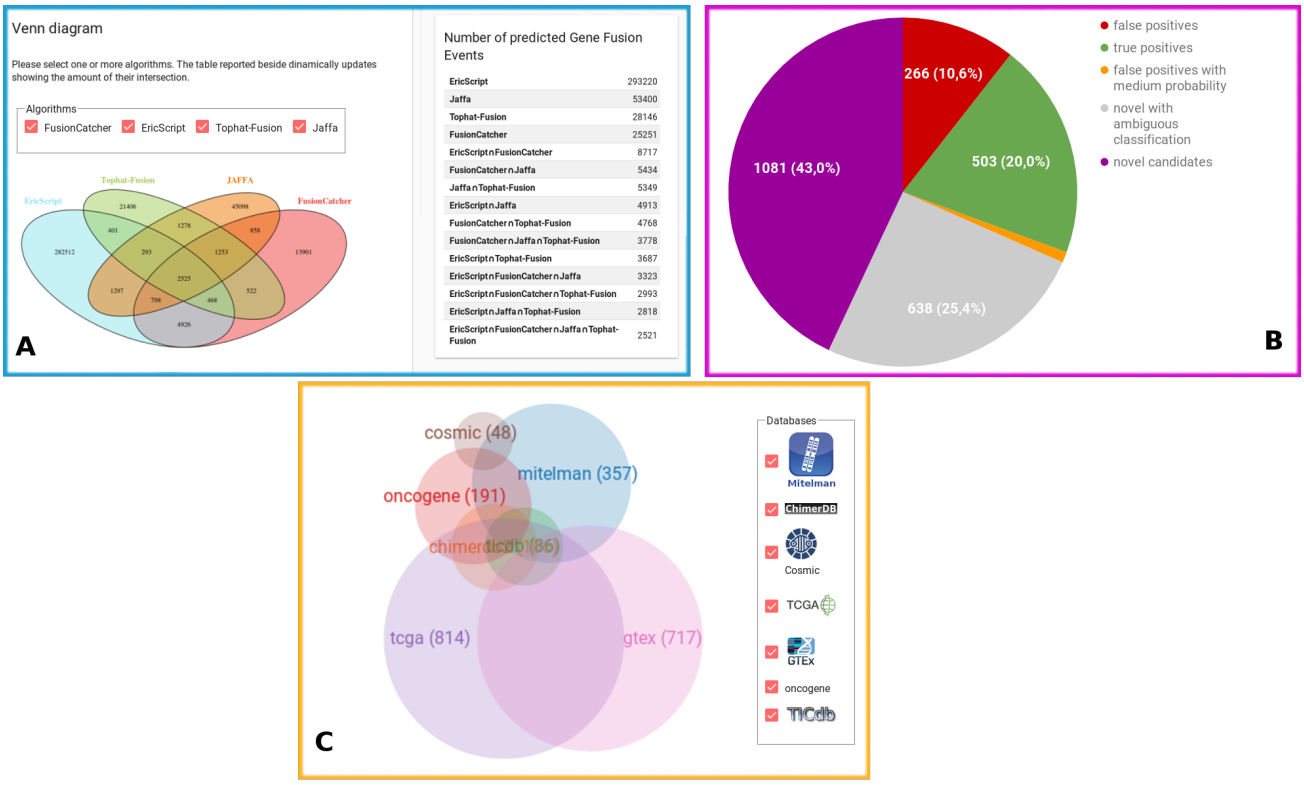

Figure 1. a) Venn diagram showing the intersection of the pGFEs identified by the four algorithms. b) Distribution of pGFEs in the Consensus Call-set: 43% (purple) of the CCS has not been previously described in any other database or scientific publication; 10% (red) and 20% (green) of the CCS have been reported in databases from healthy/tumoral samples thus representing the false/true positive subset of our analysis; 1% of the CCS (orange) reports tags which classify the pGFE as a false positive couple with medium probability; 25% (grey) of the results represent novel pGFEs tagged with values which classify them as both false and true positives. c) Venn diagram showing the intersection between the LiGeA CCS and other databases.

Figure 2. An overview of LiGeA portal. a) A 'Search by Cell line' example and the corresponding output; b) An overview of the input dataset; c) A circos diagram showing the graphical outcome of a 'Query by cell line' and the corresponding related table; d) An extract from the 'Download' web page.

terest, both the diagram and the table refresh thus showing the resulting number of intersections.

- **Search:** This utility allows several searching options to browse and mine genomic-fusion events stored in LiGeA portal (see table 2 for an overview). All the resulting outputs are sorted by the number of algorithms supporting the fusion events, thus showing on the top of the table the most robust set of results. As additional feature, when specifying the features of interest, it is also possible to choose the minimum number of predicting algorithms. Search results are presented in the form of a paginated table containing those fusion events which satisfy the query parameters and data can also be downloaded in tabular format. Furthermore, by clicking on a given fusion ID, it is possible to access the event-specific page in which relevant information is presented in greater detail (e.g., involved cell line, disease, genes as well as links to external databases and resources). Two out of nine of the query forms ('search by fusion information' and 'search by virus') are specific annotations derived FC algorithm. Here is a short description of the provided searching utilities.
- 'Search by Disease': In this section, all the cell lines derived from the same disease have been grouped together. In this way, it is possible to navigate the gene fusions putatively causing specific malignancies. The number of the cell lines constituting the queried subset is shown besides the pathology name.
- 'Search by Cell Line': This module allows to navigate the database by indicating a specific cell line name. It is possible to tune the results by showing only the novel predictions not yet described in any other database or publication (Fig.2A).
- 'Search by Chromosome': This query can be performed by inserting one or two chromosomes involved in the fusion event. The cell line name can be either indicated or not.
- 'Search by Gene': the user can select up to two gene names (Gene Symbol or ENSEMBL ID) and the 'cell line' form can be either selected or not. The genes reported in the query form are black if they are involved in pGFE and gray if they are not.
- 'Search by Transcript': Since the same gene can give rise to different transcripts, it could be reasonable to query which of the transcripts produced by a specific gene are affected by a fusion event. This kind of query can be satisfied by inserting the Ensembl Transcript (ENST) IDs in the specific form.
- 'Search by Exon': Some of the queries allow to go much more into molecular detail. This search can be done by inserting one or two exon IDs involved in the fusion event. The cell line name can be either indicated or not. In this way it is possible to highlight the specific exons which turn out to be fused in the final result.
- 'Search by Fusion information': The pGFEs may have different predicted effects. Indeed, depending on the location of the chromosomal break points, the resulting protein may be in-frame, out-of frame, truncated and so on. Since the selectable values present in the fusion information form are specific of FC algorithm, the result of this query returns a table without ES,JA and TF data. We suggest to view the FC manual in order to obtain a full description of all of the tags.
- 'Search by Algorithm': this type of query is suitable for users who wish to navigate the outputs from specific softwares, choosing them individually or in combination. Indeed, it is known that some kind of fusions, such as those involving immunoglobulins, can be detected by specific softwares 37.
- 'Search by Viruses': Another useful information retrievable from the database regards virus sequence integration into the host genome. This search utility is virus-centered since it is possible to indicate or not the host cell line name. It is possible to select the virus name of interest (whether using GI ID or NC ID). Furthermore, a clickable link redirecting to the virus genome is also shown on the right of the table.

- **Statistics:** this section allows a visual inspection of the results. The four sub-menus are organized as follows:
  - 'Cell Line Statistics': by choosing the Cell Line of interest, the resulting circular diagram shows all the chromosome couples involved in GFE predicted by at least two algorithms. The table on the right summarizes the resulting couples of the genes and chromosomes (Fig. 2C).
  - 'Chromosome Statistics': this page reports a dynamical pie-chart showing the number of fusion events per human chromosome; by clicking on each slice of the pie, the related table automatically updates showing a chromosome summary statistics. Furthermore, information about the number of inter- and intra-chromosomal rearrangements detected by each algorithm is also reported.
  - 'Disease Statistics': The 'Fusion Statistics' pie-chart was produced by grouping together the cell lines derived from the same human pathology thus showing the total number of fusion events normalized by the number of cell lines composing a specific disease. The 'Virus statistics panel' shows the frequency of exogenous virus integration per human malignancy.
  - 'Gene Statistics': A word cloud diagram showing the most recurring pGFEs supported by three methods.
  - 'Database Statistics': This sub-section is composed by four panels, the first regarding data in the CCS (Fig. 1B), the others relating only to FC and JA results. In this page it is possible to get information about the number of pGFEs found in known databases (visualized as interactive Venn diagrams and tabular fashion) and the distribution of predicted effects (histogram view).

- **Dataset:** This page is a description of the input dataset used for the analysis. Among the above 1000 samples available at the Broad institute portal 15, we downloaded 935 PE RNA-seq samples in fastq format. The SE samples have been discarded since the used softwares required it. The histogram in this section shows the number of the different cell lines derived from the same diseases (Fig. 2B). Furthermore, starting from this section, it is possible to access to web pages resuming cell-line specific details (e.g. COSMIC ID, drug resistance, human disease among others) .

- **Downloads:** From this panel it is possible to download all the processed data described within this article (Fig. 2D). Some of the files ('Summary information' and 'Viruses information') are specific products of FusionCatcher algorithm.

Availability and Requirements

- **Project name**: LiGeA: a comprehensive database of human gene fusion events
- **RRID**: SCR_015940
- **Project home page**: http://hpc-bioinformatics.cineca.it/fusion (GitHub project: https://github.com/tflati/fusion)
- **Operating system(s)**: Any
- **Programming language**: Python, JavaScript+HTML+CSS
- **Other requirements**: Django 1.10.5, Python 2.7.12, AngularJS 1.5.11
- **License**: GNU GPLv3

Declarations

List of abbreviations

LiGeA: cancer cell LInes GEne-fusions portAl; pGFE: predicted Gene Fusion Event; NGS: Next Generation Sequencing; TCGA: Tumor Cancer Genome Atlas; SRA : Sequence Read Archive; APL: acute promielocytic leukemia; CCS: Consensus Call-Set; FC: FusionCatcher; ES: EricScript; TF: Tophat-Fusion; JA: JAFFA.

Table 2. Example of possible queries on LiGeA portal

Search by	Question	Query
Disease	'what are the gene fusion events present in stomach adenocarcinoma cell lines?'	Select 'stomach adenocarcinoma' under 'disease' menu
Cell Line	'what are the novel pGFEs affecting RH30(Sarcoma) cell line?'	Select 'RH30' under the cell line menu and check the box 'show only novel results'
Chromosome	'what are the most suitable fusion partners for chromosome 8?'	Select 'Chr8' either under the '5' Chromosome' or under the '3' Chromosome' tab and leave blank the other forms
Gene	'how many human cell lines show the PML-RARA fusion event?'	Select 'PML' under the '5' gene menu'; Select 'RARA' from the '3' gene menu'; leave blank the 'Cell Line' query form;
Fusion information	'what are all the in-frame pGFEs in Jurkat cell line?'	select 'Jurkat' under 'Cell line' menu; Select 'in-frame' under 'predicted effect menu
Fusion information	'what are the known GFEs predicted to be in-frame in Jurkat cell line?'	Select 'Jurkat' under 'Cell line' menu; Select 'in-frame' under 'predicted effect menu; select 'known' under 'Fusion description' menu
Algorithm	'show only those GFEs supported by FC and TF in RH30 cell line'	Select 'RH30' under 'Cell Line' query form and check the boxes relative to FC and TF
Viruses	'which cell lines are most affected by Hepatitis C virus genome integration?'	Select 'Hepatitis C virus' under 'Virus' query form and let blank the 'Cell line' query form

Consent for publication

'Not applicable'

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was supported by ELIXIR-IIB, CINECA and Regione Lombardia. In particular, Enrico Garattini was supported by the "Fondazione Italo Monzino" and AIRC (Award ID 17058) fundings. Maddalena Fratelli was supported by "ELIXIR-IIB-Cineca". Silvia Gioiosa was funded by "ELIXIR-IIB", program name "Efficient implemenation and distribution of HPC bioinformatics resources for Elixir scientific community", Award Number: 08/AR/2016-IBBE-BA. Tiziano Flati was funded by "ELIXIR-IIB", program name "Efficient allocation of HPC bioinformatics resources through a federation of Galaxy web-based infrastructures", Award Number: 05/AR/2016-IBBE-BA.

Author's Contributions

TC and MF conceived and designed the work. All authors analyzed, interpreted data, wrote the manuscript and approved the final manuscript.

Acknowledgements

We acknowledge Andrea Micco for his useful tests on the first prototype of the system. We acknowledge the CINECA and the Regione Lombardia award under the LISA initiative 2016-2018, for the availability of high performance computing resources and support. In particular the ELIXIR-ITA HPC@CINECA and ELIXIR-IIB HPC@CINECA initiatives for providing HPC resources to our project.

Table 1. State of the art of databases reporting gene fusions

Database Name	Short Description
Tumor Fusion Gene Data Portal 7	A collection of fusion genes in the Tumor Cancer Genome Atlas (TCGA) samples.
TICdb 8	A collection of 1,374 fusion sequences extracted either from public databases or from published papers (last update: 2013).
chimerDB3.0 9	A catalog of fusion genes encompassing analysis of TCGA data and manual curations from literature.
COSMIC Cell Lines 10	Gene fusions are manually curated from peer reviewed publications. Currently COSMIC includes information on fusions involved in solid tumors but not yet leukemias and lymphomas.
Mitelman 1	Reports hundreds of gene fusions associated with clinical reports but does not contain sequence data.
ChiTaRs-3.1 11	A collection of 34,922 chimeric transcripts identified by Expressed Sequence Tags (ESTs) and mRNAs from the GenBank, ChimerDB, dbCRID, TICdb and the Mitelman collection of cancer fusions for several organisms.
FusionCancer 12	591 samples, both single-end and paired-end RNA-seq, published on SRA database 13 between 2008 and 2014 covering 15 kinds of human cancers .
ONGene 14	Literature-derived database of oncogenes.

References

[1] Mitelman F, Johansson B, Mertens F. "The impact of translocations and gene fusions on cancer causation". Nat Rav Cancer 2007;7(4):233 – 245.

[2] Mertens F, Johansson B, Fioretos T, Mitelman F. The emerging complexity of gene fusions in cancer. Nat Rev Cancer 2015;15(6).

[3] Serratì S, De Summa S, Pilato B, Petriella D, Lacalamita R, Tommasi S, et al. Next-generation sequencing: advances and applications in cancer diagnosis. OncoTargets and Therapy 2016;9:7355–7365.

[4] Borrow J, Goddard A, Sheer D, Solomon E. Molecular analysis of acute promyelocytic leukemia breakpoint cluster region on chromosome 17. Science 1990;249(4976):1577–1580. http://science.sciencemag.org/content/249/4976/1577.

[5] Nervi C, Ferrara FF, Fanelli M, Rippo MR, Tomassini B, Ferrucci PF, et al. Caspases Mediate Retinoic Acid–Induced Degradation of the Acute Promyelocytic Leukemia PML/RARα Fusion Protein. Blood 1998;92(7):2244–2251. http://www.bloodjournal.org/content/92/7/2244.

[6] Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, et al. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences of the United States of America 2009;110(49):19872–19877.

[7] TCGA, Tumor Fusion Gene Data Portal @ONLINE;. http://54.84.12.177/PanCanFusV2//.

[8] Novo F, de Mendíbil I, Vizmanos J. TICdb: a collection of gene-mapped translocation breakpoints in cancer. BMC Genomics 2007;8(33). http://www.unav.es/genetica/TICdb/.

[9] Lee M, Lee K, Yu N, Jang I, Choi I, Kim P, et al. ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining. Nucleic Acids Research 2017;45(D1):D784–D789. http://dx.doi.org/10.1093/nar/gkw1083.

[10] COSMIC, COSMIC Database-Wellcome Trust Sanger Institute @ONLINE; 2017. http://cancer.sanger.ac.uk/cell_lines.

[11] Gorohovski A, Tagore S, Palande V, Malka A, Raviv-Shay D, Frenkel-Morgenstern M. ChiTaRS-3.1—the enhanced chimeric transcripts and RNA-seq database matched with protein–protein interactions. Nucleic Acids Research 2017;45(D1):D790–D795. http://chitars.md.biu.ac.il/index.html.

[12] Wang Y, Wu N, Liu J, Wu Z, Dong D. FusionCancer: A database of cancer fusion genes derived from RNA-seq data 2015 12;10:131. http://donglab.ecnu.edu.cn/databases/FusionCancer/.

[13] SRA, Sequence Read Archive - SRA @ONLINE;. http://www.ncbi.nlm.nih.gov/sra.

[14] Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. Journal of Genetics and Genomics 2017;44(2):119 – 121. http://ongene.bioinfo-minzhao.org/.

[15] CCLE, Broad Institute portal - CCLE Repository;. https://portals.broadinstitute.org/ccle/home.

[16] Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014;http://www.biorxiv.org/content/early/2014/11/19/011650.

[17] Benelli M, Pescucci C, Marseglia G, Severgnini M, Torricelli F, Magi A. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. Bioinformatics 2012;28(24):3232–3239. http://dx.doi.org/10.1093/bioinformatics/bts617.

[18] Daehwan K, Salzberg S. TopHat-Fusion: An Algorithm for Discovery of Novel Fusion Transcripts. Genome Biology 2011;12(8).

[19] Davidson NM, Majewski IJ, Oshlack A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. Genome Medicine 2015 May;7(1):43. https://doi.org/10.1186/s13073-015-0167-x.

[20] Kumar S, Vo AD, Qin F, Li H. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. Nature Scientific Reports 2016;6.

[21] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 2009 Mar;10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25.

[22] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature methods 2012;9(4):357–359.

[23] Kent W. BLAT—The BLAST-Like Alignment Tool. Genome Research 2002;12(4):656–664.

[24] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21. http://dx.doi.org/10.1093/bioinformatics/bts635.

[25] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25(9):1105–1111. +http://dx.doi.org/10.1093/bioinformatics/btp120.

[26] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–2120. +http://dx.doi.org/10.1093/bioinformatics/btu170.

[27] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.

[28] Sadedin SP, Pope B, Oshlack A. Bpipe: a tool for running and managing bioinformatics pipelines. Bioinformatics 2012;28(11):1525–1526. +http://dx.doi.org/10.1093/bioinformatics/bts167.

[29] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2008, http://www.R-project.org, ISBN 3-900051-07-0.

[30] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754–1760. +http://dx.doi.org/10.1093/bioinformatics/btp324.

[31] Ensembl DataBase @ONLINE;. https://docs.google.com/uc?id=0B9s__vuJPvIiUGt1SnFMZFg4TlE&export=download.

[32] Shugay M, Ortiz de Mendíbil I, Vizmanos JL, Novo FJ. Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics 2013;29(20):2539–2546. +http://dx.doi.org/10.1093/bioinformatics/btt445.

[33] Gioiosa S, Bolis M, Flati T, Massini A, Garattini E, Chillemi G, et al. Supporting data for 'Massive NGS Data Analysis Reveals Hundreds Of Potential Novel Gene Fusions in Human Cell Lines'. GigaScience Database 2018;http://dx.doi.org/10.5524/100442.

[34] COSMIC, COSMIC Database-Wellcome Trust Sanger Institute @ONLINE; 2017. http://cancer.sanger.ac.uk/cosmic.

[35] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A CENSUS OF HUMAN CANCER GENES. Nature reviews Cancer 2004;4(3):177–183.

[36] COSMIC, COSMIC Gene Census - Wellcome Trust Sanger Institute @ONLINE; 2017. http://cancer.sanger.ac.uk/census.

[37] Reshmi SC, Harvey RC, Roberts KG, Stonerock E, Smith A, Jenkins H, et al. Targetable kinase gene fusions in high-risk B-ALL: a study from the Children's Oncology Group. Blood 2017;129(25):3352–3361. http://www.bloodjournal.org/content/129/25/3352.