ChiTaH: a fast and accurate tool for identifying known human chimeric sequences from high-throughput sequencing data

Rajesh Detroja, Alessandro Gorohovski, Olawumi Giwa, Gideon Baum and Milana Frenkel-Morgenstern

Cancer Genomics and BioComputing of Complex Diseases Lab, Azrieli Faculty of Medicine, Bar-Ilan University, Safed 1311502, Israel

Received June 23, 2021; Revised October 21, 2021; Editorial Decision November 08, 2021; Accepted November 22, 2021

ABSTRACT

Fusion genes or chimeras typically comprise sequences from two different genes. The chimeric RNAs of such joined sequences often serve as cancer drivers. Identifying such driver fusions in a given cancer or complex disease is important for diagnosis and treatment. The advent of next-generation sequencing technologies, such as DNA-Seq or RNA-Seq, together with the development of suitable computational tools, has made the global identification of chimeras in tumors possible. However, the testing of over 20 computational methods showed these to be limited in terms of chimera prediction sensitivity, specificity, and accurate quantification of junction reads. These shortcomings motivated us to develop the first ‘reference-based’ approach termed ChiTaH (Chimeric Transcripts from High–throughput sequencing data). ChiTaH uses 43,466 non–redundant known human chimeras as a reference database to map sequencing reads and to accurately identify chimeric reads. We benchmarked ChiTaH and four other methods to identify human chimeras, leveraging both simulated and real sequencing datasets. ChiTaH was found to be the most accurate and fastest method for identifying known human chimeras from simulated and sequencing datasets. Moreover, especially ChiTaH uncovered heterogeneity of the BCR−ABL1 chimera in both bulk and single-cells of the K-562 cell line, which was confirmed experimentally.

INTRODUCTION

Gene-gene fusions (or chimeric RNAs generated at the RNA level) typically comprise sequences from two different genes. Fused genes can be generated by different mechanisms, including chromosomal translocations, transcriptional errors, or by cis- or trans-splicing (1–5). The chimeric RNAs of such joined sequences often serve as cancer drivers, as exemplified by BCR−ABL1, found in ~95% of chronic myelogenous leukemia (CML) cases (6), by TMPRSS2−ETS, found in ~50% of instances of prostate cancer (7), and by DNAJB1−PRKACA, the hallmark and likely driver of fibrolamellar carcinoma (8). Identifying such driver fusions in a given cancer or complex disease is important for diagnosis and treatment. For example, tyrosine–kinase inhibitors are highly effective in treating patients suffering from CML and other cancers harboring kinase fusions (9–12). Despite the importance of chimeric RNAs in various cancers, the current terminology used to describe gene-gene fusions can be inconsistent, possibly due to the field’s infancy. Presently, numerous terms have been used to describe chimeric RNAs, such as transcription-mediated fusions, fusion genes, conjoined genes, complex genes, chimeras, spanning genes, hybrid genes, and fusion transcripts. In this article, we define chimeric RNAs and fusion genes as chimeras.

Cytogenetic analysis led to the discovery of BCR−ABL1 as the first chimera in 1973 (13). Since, dozens of chimeras have been identified in hematologic cancers. Over the past decade, the advent of high–throughput, low–cost, and sophisticated next-generation sequencing (NGS) technologies, such as DNA-Seq or RNA-Seq, together with improved computational power, has made global identification of chimeras in solid tumors, including sarcoma, carcinoma, and tumors of the central nervous system, possible (14). Today, over 20 computational tools for identifying chimeras using high–throughput sequencing data are available. The common feature of these de novo chimera identification methods is the use of discordant and/or split reads to recognize exon-exon junction sites (15–34). At the same time, the various methods vary in terms of the read alignment tool employed, the version of the human reference genome consulted, the gene annotations used, the criteria
selected for filtering candidate chimeras, and likely, false positives. Furthermore, these tools were found to be limited in terms of prediction sensitivity and specificity, accurate quantification of junction reads, installation complexity, execution time and robustness when utilizing DNA-Seq or RNA-Seq and paired-end or single-end sequencing data (35–37).

These shortcomings motivated us to develop a first ‘reference-based’ approach, termed ChiTaH (Chimeric Transcripts from High-throughput sequencing data), for identifying chimeras from both high-throughput DNA-Seq and RNA-Seq sequencing data. Since the introduction of NGS, thousands of human chimeras have been identified and published at NCBI and in the literature (38). Accordingly, we processed 10,000,000 human EST/mRNA sequences from NCBI GenBank and identified 548,262 human chimeric transcripts and collected them into the latest version of our extended ChiTaRS 5.0 database (39). These chimeric transcripts were then used to assemble a reference database of chimeras presenting non-redundant unique junction sequences that included 43,466 humans chimeric RNAs. ChiTaH uses these 43,466 chimeras as a reference database of known human chimeras to map DNA-Seq or RNA-Seq sequencing reads to accurately identify chimeric reads. We subsequently evaluated the performance of ChiTaH and four best chimer detection methods, namely, EricScript (19), STAR-Fusion (16), JAFFA (15), and FusionCatcher (20). We assessed each method on simulated and real sequencing datasets by evaluating various parameters, such as sensitivity and specificity to detect candidate chimeras, quantification of detected junction reads, total time required, and RAM (random-access memory) consumed to complete the analysis. Moreover, we analyzed single-cell RNA-Seq sequencing data collected from the K562 cell line to assess the potential of ChiTaH to determine BCR-ABL1 chimera heterogeneity and expression at the single cell level. We experimentally confirmed the heterogeneity of BCR-ABL1 in the K562 cell line. Finally, we analyzed RNA-Seq samples of human cancer cell lines from the CCLE (Cancer Cell Line Encyclopedia) to build a catalog of cancer-specific chimeras. We also analyzed, RNA-Seq samples of normal human tissues from ArrayExpress archive of EBI (European Bioinformatics Institute) to build a catalog of normal tissue-specific chimeras, using ChiTaH.

MATERIALS AND METHODS

Collection of known human chimeras

A total of 10,000,000 human EST/mRNA sequences were retrieved from the NCBI GenBank database and aligned to a human reference genome (hg38) using the BLAT program (40). EST/mRNA sequences were considered as human chimeras if their first and second sequence tracts shared a minimum identity of 95% and if these two sequences could not be mapped linearly to the human reference genome. In addition, the total length of the two connected mapped nucleotide sequence tracts had to be more than one-third of the length of the original EST/mRNA sequence. Original EST/mRNA sequences with lengths shorter than 21 bp and longer than 75,000 bp were discarded. A total of 548,262 human chimeras were thus identified and included in the latest version of our extended ChiTaRS 5.0 database (39).

Pre-processing of the human chimeras was performed to filter out low-quality sequences. As such, chimeras were filtered out if the identities of the gene1 and gene2 sequences to the human reference genome (hg38) were <95%, or if the chimeras presented an ‘intron–intron’ junction type, containing only non-coding sequences in the parental genes. Isoform chimeras with identical parental genes that were identically orientationally were removed, as were chimeras with short-length parental genes of less than 50 bp. Following such screening, a total of 74,223 high-quality human chimeras remained (Figure 1A).

Non-redundant known human chimeras

A total of 74,223 (100%) high-quality chimeras were used to prepare human chimeras with non-redundant unique junction sequences. To calculate the junction region of a chimera, the overlap length between the gene1 and gene2 sequences was calculated as follows:

Overlap start = IF (S2 < E1 then S2 else E1)
Overlap end = IF (S2 < E1 then E1 else S2)
Overlap length = (Overlap end − Overlap start) + 1

Chimeras with overlap lengths of more than 40 bp were filtered out (Figure 1B). The remaining chimeras were used for the junction length calculations as follows:

Junction start = IF (overlap start − 15 >
= S1 then overlap start − 15 else S1)
Junction end = IF (overlap end + 15 <
= E2 then overlap end + 15 else E2)
Junction length = (Junction end − Junction start) + 1

To identify non-redundant unique junction regions, stand-alone BLASTn (41) of 65,761 junction sequences were performed against the NCBI curated databases of human genomic sequences (ref_euk_rep_genomes; taxids:9606) and human transcripts (refseq_rna; taxids:9606), a human reference genome (hg38), and a human reference transcriptome (hg38:cDNA) with 95% minimum identity and 100% query coverage. In total, 15187 (20.46%) chimeras with identical junction sequences as human sequences were filtered out (Figure 1D). The remaining junction sequences were subjected to analysis using dedupe.sh (https://github.com/BioInfoTools/BBMap) to detect duplicate junction sequences by pairwise alignment. In total, 4280 (5.76%) chimeras with duplicate junction sequences were excluded (Figure 1E). The remaining junction sequences were used to calculate the percentage of A, T and N nucleotides, using the faCount (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/facount) tool. Chimeras indicative of the possible presence of a poly-A tail or ambiguous nucleotides containing a high percentage of A, T, and N nucleotides at junction sequences were filtered out (Figure 1F). Thus, 43,466 (58.56%) high-quality, non-redundant human chimeras were obtained.
ChiTaH uses 43,466 non-redundant unique human chimeras, a human reference genome (hg38), and a human reference transcriptome (hg38:cDNA) as a combined reference to map DNA-Seq or RNA-Seq sequencing reads using the bowtie2 aligner with a local alignment approach (42). ChiTaH subsequently calculates chimeric reads mapped at non-redundant unique junction regions of chimeras using bedtools coverage (43). Then, ChiTaH calculates final candidate chimeras and provides the output of all identified chimeras across samples with junction read counts offered in a single matrix table (Figure 2C).

Comparative assessment of chimera detection methods

We sought to comparatively evaluate the performances of ChiTaH and the current best chimera detection methods. Over the past decade, over 20 computational tools for identifying chimeras using high-throughput sequencing data have become available. Earlier studies performed comprehensive evaluation of many of these methods (35–37). Based on their conclusions, we downloaded and installed four tools, i.e., EricScript (19), STAR-Fusion (16), JAFFA (15), and FusionCatcher (20), for benchmarking (Table 1). All four methods and ChiTaH were run on a simulated dataset and bulk RNA-Seq sequencing data at default, using recommended parameters for each method, 25 CPUs, and 1000 Gb of RAM. The sensitivity and specificity of each tool on the simulated dataset were assessed using the following criteria:

1. Sensitivity (%) = (TP/TF) × 100%
2. Positive Predictive Value (PPV) or Specificity (%) = (TP/(TP + FP)) × 100%

TP: True positive, correctly identified known candidate chimeras
TF: Total known chimeras in the simulated dataset
FP: False-positive chimeras identified

For each run of the simulated and bulk RNA-Seq sequencing datasets, the computational RAM used in gigabytes (Gb) and time consumed in minutes to complete the analysis were also calculated.

Simulated dataset

The simulated dataset was generated for comparative assessment analysis of the different chimera detection methods. Sequences of 100 known chimeras in humans were used to assess the sensitivity of each method. The sequence of each of the 100 known chimeras is available at NCBI and contains unique parental genes and non-redundant unique junction sequences. About 95% of known chimeras contain exons in both parental genes, and can thus be translated into protein (Supplementary Figure S1). Human CDS sequences with a minimum of 2000 bp were obtained from Ensemble and merged with sequences of 100 known chimeras to generate an artificial mRNA assembly ~60 Mb in size, containing a total of 17 163 sequences. This artificial mRNA assembly was used to generate simulated sequencing data using BBMap reformat.sh (https://github.com/BioInfoTools/BBMap). In to-
Figure 2. Workflow of ChiTaH. (A) Paired-end and single-end NGS sequencing reads as input to ChiTaH. (B) Mapping of reads using Bowtie2 against a reference database comprising a human genome, a human transcriptome and 43,466 non–redundant known human chimeras. (C) Prediction of candidate chimeras based on calculation of junction reads.

Table 1. Four top-level chimera detection methods used for benchmarking

<table>
<thead>
<tr>
<th>Tool name</th>
<th>Read type supported</th>
<th>Reference</th>
<th>Version</th>
<th>Sequencing data type</th>
<th>Interface</th>
<th>Alignment algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>EricScript</td>
<td>Paired-end</td>
<td>(19)</td>
<td>0.5.5</td>
<td>RNA-Seq</td>
<td>Stand-alone</td>
<td>BWA (44), BLAT (40)</td>
</tr>
<tr>
<td>STAR-Fusion</td>
<td>Paired-end &amp; single-end</td>
<td>(16)</td>
<td>1.9.1</td>
<td>RNA-Seq</td>
<td>Stand-alone</td>
<td>STAR (45)</td>
</tr>
<tr>
<td>JAFFA</td>
<td>Paired-end &amp; single-end</td>
<td>(15)</td>
<td>2.00</td>
<td>RNA-Seq</td>
<td>Stand-alone</td>
<td>bowtie2 (42), BLAT (40)</td>
</tr>
<tr>
<td>FusionCatcher</td>
<td>Paired-end &amp; single-end (150 bp)</td>
<td>(20)</td>
<td>1.20</td>
<td>RNA-Seq</td>
<td>Stand-alone</td>
<td>STAR (45), BLAT (40), bowtie2 (42)</td>
</tr>
</tbody>
</table>

Bulk RNA-Seq sequencing dataset

A bulk RNA-Seq sequencing dataset was prepared to evaluate parameters such as total detected junction reads, and time and RAM needed to complete the run of real sequencing data. Three bulk RNA-Seq sequencing samples from the K-562 cell line were downloaded from the NCBI SRA database. This source, positive for the known fusion gene BCR–ABL1, contains about 53, 50 and 54 million paired-end reads of 101 bp for samples SRR9032085, SRR9032086 and SRR9032088, respectively.

Single-cell RNA-Seq sequencing dataset

To assess the potential of ChiTaH to determine the heterogeneity of BCR–ABL1 chimeras and their expression at the single-cell level, single-cell sequences from the K-562 cell line were downloaded from NCBI BioProject accession PRJNA430491. To generate the dataset, the Smart-Seq2 design and BGISEQ-500 platform were used to sequence a total of 81 and 249 single cells as paired-end and single-end reads, respectively.

Dataset of cancer cell lines and normal tissues

To assemble a catalog of cancer-specific fusion genes using ChiTaH, 934 RNA-Seq samples of human cancer cell lines from the CCLE (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2770/Downloads) and 199 RNA-Seq samples of 32 normal human tissues from the ArrayExpress archive of EBI (https://www.ebi.ac.uk/gxa/experiments/E-MTAB-2836/Downloads) were analyzed.

Hardware, software and statistical analysis

Development and comparative analysis were performed on the RedHat Enterprise Linux 7.4 server of an Intel(R) Xeon(R) CPU E5-2620 v2 server with 25 CPUs and 1000
ChiTaH: a reference-based approach for detection of human chimeras

After pre-processing human EST and mRNA sequences from NCBI, a total of 43,466 high-quality, non-redundant human chimeras were curated. A total of 38,550 (88.69%) chimeras contained 1–10 bp of overlap length, followed by 3652 (8.40%) chimeras containing 11–20 bp overlap length, 855 (1.96%) chimeras containing 21–30 bp overlap length and 409 (0.94%) chimeras containing 31–40 bp overlap length were identified. Each chimera junction length was calculated by adding 15 bp to both sides of the overlap region. Hence, junction length is correlated with overlap length of the chimera. This yielded 38,550 (88.69%) chimeras containing 31–40 bp of junction length, followed by 3652 (8.40%) chimeras containing 41–50 bp of junction length, 855 (1.96%) chimeras containing 51–60 bp of junction length, and 409 (0.94%) chimeras containing 61–70 bp of junction length. Moreover, 25,252 (58%) chimeras contain exons in both parental gene sequences, with the potential of translation into a complete chimeric protein. In contrast, a total of 9004 (21%) and 9210 (21%) chimeras contain intron-exon and exon-intron combinations in their parental gene sequences, respectively, and thus have the potential of translation into a partially chimeric protein (Supplementary Figure S4).

ChiTaH uses these chimeras to identify chimeric reads from the DNA-Seq or RNA-Seq sequencing data. Specifically, ChiTaH uses only non-redundant junction regions to identify chimeric reads or chimeras, which makes ChiTaH more accurate and robust when considering paired-end or single-end sequencing reads of variable lengths of more than 35 bp.

Comparative assessment of chimera detection methods using a simulated dataset

ChiTaH and four current best methods, namely, EricScript (19), STAR-Fusion (16), JAFFA (15) and FusionCatcher (20), were run on simulated datasets. For each method, default pre-defined alignment and analysis parameters were used. Using simulated datasets, the performance of each method was evaluated in terms of various parameters, such as sensitivity, specificity, quantification of junction reads, and total time and RAM utilized to complete the analysis. Sensitivity in identifying the 100 known chimeras was first assessed. Sensitivity was calculated by considering the number of known chimeras identified by each method from the total of 100 known chimeras in the simulated datasets (Supplementary Figure S2C, D). The average sensitivity of ChiTaH was found to be 100% for identifying known chimeras from paired-end datasets, followed by FusionCatcher and EricScript, with average sensitivities of 89.75% and 80.75%, respectively. In identifying known chimeras from single-end datasets, the average sensitivity of ChiTaH was found to be 100%, followed by FusionCatcher and JAFFA, with average sensitivities of 89% and 81.50%, respectively (S4 File).

Specificity or positive predictive value (PPV) was next assessed to identify known chimeras. Specificity (%) was calculated by considering both the number of known chimeras and the number of false-positive chimeras identified by each method from simulated datasets (Supplementary Figure S2E, F). The average specificity of ChiTaH in identifying chimeras from paired-end datasets was found to be 100%, followed by STAR-Fusion and FusionCatcher, with average specificities of 92.40% and 88.02%, respectively. In identifying chimeras from single-end datasets, the average specificity of ChiTaH was found to be 100%, followed by STAR-Fusion and FusionCatcher, with average specificities of 92.34% and 84.76%, respectively.

The ability to accurately estimate the expression or abundance of identified known chimeras by counting junction reads was next determined. To calculate identified junction reads by each method, a total of 45 commonly identified known chimeras by all five methods in paired-end and/or single-end simulated datasets were listed. Next, a total number of average simulated junctions reads generated for these 45 chimeras was calculated. Thus, on average, 25.91 and 27.88 simulated junctions read were generated for 45 chimeras in paired-end and single-end datasets, respectively. These were considered as control (Min-Expected). From paired-end datasets, ChiTaH identified an average of 28.41 junction reads, followed by STAR-Fusion and EricScript, which identified an average of 25.76 and 22.66 junction reads, respectively. In single-end datasets, ChiTaH
identifying an average of 29.92 junction reads, followed by STAR-Fusion and FusionCatcher, which identified an average of 24.54 and 23.27 junction reads, respectively (Figure 3E, F). Moreover, it was noted that ChiTaH identified a few more junction reads than expected, possibly because the sensitivity of ChiTaH also allows for the recognition of junction reads presenting mutations.

The total time in minutes used by each method to complete runs on the simulated datasets was next determined. START-Fusion utilized an average of 9.5 minutes to complete runs of paired-end datasets followed by ChiTaH and FusionCatcher which utilized an average of 11.32 and 51.96 minutes, respectively. Where, to complete runs of single-end datasets ChiTaH utilized an average of 2.96 minutes followed by STAR-Fusion and JAFFA which utilized an average of 7.99 and 224.24 minutes, respectively (Figure 4A, B). Finally, the total RAM consumed in Gb by each method to complete analysis of the run-on simulated datasets was assessed. ChiTaH consumed an average of 4.52 Gb of RAM to complete runs of paired-end datasets followed by EricScript and JAFFA which consumed an average of 18.87 and 20.82 Gb of RAM. To complete runs of single-end datasets, ChiTaH utilized an average of 4.32 Gb of RAM, followed by JAFFA and FusionCatcher, which consumed 9.53 Gb and 34.19 Gb of RAM, respectively (Figure 4C, D). Altogether, comparative benchmarking of ChiTaH and the four other methods on simulated datasets found ChiTaH to be the most accurate and fastest method for identifying known human chimeras, followed by STAR-Fusion.

Comparative assessment of chimera detection methods using real sequencing datasets

After benchmarking all five methods on simulated datasets, correlation of the same results with real sequencing datasets was tested. For this, three bulk RNA-Seq samples from the myelogenous leukemia K-562 cell line, namely, SRR9032085, SRR9032086 and SRR9032088, were downloaded from SRA. In these sequencing datasets, the true number of chimeras was not known. As such, parameters such as sensitivity and specificity could not be assessed. Nevertheless, the K-562 cell line-derived samples are positive for the chimera BCR-ABL1, allowing for evaluation of the quantification of BCR-ABL1 junction reads, the total time required, and RAM utilized. Each method was run on the K-562 cell line sequencing datasets using the same server, 25 CPUs, and 1000 Gb of RAM. The total number of junctions reads detected by each method was first calculated. ChiTaH detected an average of 119.33 BCR-ABL1 junction reads, followed by STAR-Fusion and EricScript, which detected an average of 99.33 and 83 junctions, respectively (Figure 5A).

The total time in minutes used to complete the run by each method was also determined. STAR-Fusion utilized an average of 19.97 minutes, followed by ChiTaH and JAFFA, which utilized an average of 33.86 and 171.72 minutes, respectively. Finally, the total RAM in Gb to complete the run by each method was assessed. ChiTaH consumed an average of 4.67 Gb of RAM, followed by JAFFA and EricScript, which consumed an average of 16.36 and 23.49 Gb of RAM, respectively (Figure 4E, F). Thus, again in agreement with the results obtained using the simulated datasets, ChiTaH and STAR-Fusion were found to be the most accurate and fastest method for identifying BCR-ABL1 chimeras from real sequencing datasets of K-562 cell line.

Expression of BCR-ABL1 chimeras in bulk and single cells of the K-562 cell line

BCR-ABL1 chimera was detected by all five methods from bulk RNA-Seq of the K-562 cell line. However, ChiTaH uniquely identified two different chimeric sequences of BCR-ABL1, namely FUS390 and M19695. FUS390 contains exons in both of the parental genes, with the potential of being translated into a complete chimeric protein, while M19695 contains exons in the BCR gene and introns in the ABL1 gene, with the potential of being partially translated into a chimeric protein. A total of 76, 103 and 96 FUS390 junction reads were identified from samples SRR9032085, SRR9032086 and SRR9032088, respectively. A total of 38, 25 and 20 M19695 junction reads were identified from samples SRR9032085, SRR9032086 and SRR9032088, respectively (Figure 5C). To confirm differences at the sequence level, more than 100 bp of junction sequences of FUS390 and M19695 were aligned using the EBI ClustalW (46) program. The alignment of junction sequences showed BCR gene regions to be identical between FUS390 and M19695, while ABL1 regions were unrelated. Further, in order to validate two different version of BCR-ABL1 chimeras in the K562 cell line, we performed a PCR assay designed to detect and amplify these chimeras in K-562 cell cDNA. The PCR product were of the expected sizes. The product was eluted from the gel and subjected to Sanger sequencing with the same primers used for PCR amplification. The sequencing results of both DNA strands were identical to the sequences found by ChiTaH, confirming the presence of FUS390 and M19395 chimeras in the K-562 cell line (Supplementary Figure S5).

Taken together, this result shows the existence of two different versions of BCR-ABL1 chimeras in the K-562 cell line, namely, FUS390, with significantly high expression, and M19695, with significantly low expression (P-value: 0.042) (Figure 5B). Based on the discovery of distinct BCR-ABL1 chimeras by ChiTaH in bulk RNA-Seq data of K-562 cells, it was hypothesized that these differences in the BCR-ABL1 chimera could be due to cellular heterogeneity in the K-562 cell line at the single-cell level. To validate this hypothesis, a publicly available total of 81 and 249 K-562 cell line single cells with RNA-Seq sequencing data as paired-end and single-end reads, respectively, were downloaded and analyzed using ChiTaH. With the paired-end single-cell dataset, BCR-ABL1 chimeras were found in a total of 12 (14.81%) cells, with the FUS390 chimera being found in 9 (11.11%) cells, and the M19695 chimera being found in 3 (3.70%) cells. In the single-end single-cell dataset, BCR-ABL1 chimeras were found in a total of 16 (6.41%) cells, with the FUS390 chimera being found in 11 (4.41%) cells, and the M19695 chimera being found in 5 (2%) cells. In the paired-end or single-end datasets, no cells containing both versions of the BCR-ABL1 chimera were detected. Furthermore, the expression of each chimera in the cell was normalized to RPM (reads per million) and
Figure 3. Comparative assessment of ChiTaH using simulated datasets. (A) Sensitivity (%) of the five methods tested on paired-end simulated datasets. (B) Sensitivity (%) of four methods on single-end simulated datasets. (C) PPV (%) of the five methods on paired-end simulated datasets. (D) PPV (%) of four methods on single-end simulated datasets. (E) Average junction reads detected by the five methods from paired-end simulated datasets. (F) Average junction reads detected by four methods from single-end simulated datasets.
Figure 4. Comparative assessment of ChiTaH using simulated and real datasets. (A) Total time in minutes used by the five methods to complete a run of paired-end simulated datasets. (B) Total time in minutes used by four methods to complete a run of single-end simulated datasets. (C) Total memory in Gb used by the five methods to complete a run of paired-end simulated dataset. (D) Total memory in Gb used by four methods to complete a run of single-end simulated datasets. (E) Total time in minutes used by the five methods to complete a run of the K-562 cell line sequencing datasets. (F) Total memory in Gb used by the five methods to complete a run of the K-562 cell line sequencing datasets.
Figure 5. Expression of BCR-ABL1 chimeras calculated by ChiTaH. (A) Total junction reads detected by the five methods in the K-562 cell line sequencing datasets. (B) Expression of BCR-ABL1 chimeras in the K-562 cell line bulk RNA-Seq dataset. (C) ChiTaH identified a distinct junction in BCR-ABL1 in the K-562 cell line bulk RNA-Seq dataset. (D) Expression of BCR-ABL1 chimeras in the K-562 cell line paired-end single-cell dataset. (E) Expression of BCR-ABL1 chimeras in the K-562 cell line single-end single-cell dataset.
an unpaired t-test was performed to compare the expression of FUS390 and M19695 at the single cell level. With paired-end or single-end single-cell data, no significant differences were found between the expression of FUS390 and M19695 (Figure 5D, E), which is completely distinct from what was observed with the bulk RNA-Seq data. This discrepancy could be due to the significantly lower expression of FUS390 in most of the cells. Overall, these findings support the presented hypothesis and revealed the heterogeneity of BCR-ABL1 chimeras at the single-cell level using ChiTaH. Unlike other chimera detection methods, ChiTaH identified chimeras at the bulk and single-cell levels and was quite accurate in detecting heterogeneity in known human chimeras.

**Chimera analysis of cancer cell lines and normal human tissues**

To build a catalog of cancer-specific chimeras using ChiTaH, 934 RNA-Seq samples of human cancer cell lines from the CCLE and 199 RNA-Seq samples of 32 normal human tissues from the ArrayExpress archive of EBI were analyzed. In the CCLE, 934 cell lines of 26 different cancer tissues were sequenced, including 173, 162, 54, 50 and 49 cancer samples from lung, hematopoietic/lymphoid tissues, large intestine, breast, and central nervous system, respectively (Supplementary Figure S3A). After analysis of these datasets using ChiTaH, a total of 2066 and 7673 chimeras were found in the EBI and CCLE samples, respectively. Of these, 1717 chimeras were found to be common to the EBI and CCLE lists, and were considered as normal population chimeras (S2 File) (S3 File). A total of 5956 chimeras were found to be unique to the CCLE cancer samples, and were considered as cancer-specific chimeras (S1 File). Distribution of the junction type in the 5956 chimeras showed that about 3666 (62%) chimeras contain exons in both parental genes, with the potential of being translated into a complete chimeric protein, while 1133 (19%) and 1157 (19%) chimeras contained exon-intron and intron-exon pairs in their parental genes, with the potential to be partially translated into a chimeric protein (Supplementary Figure S3B, C). A total of 3073, 3060, 2009, 1946 and 1568 chimeras were identified from the cancerous lung, hematopoietic/lymphoid tissues, large intestine, breast, and central nervous system, respectively. Moreover, a total of six distinct versions of the BCR-ABL1 chimera was found across 14 cell lines of hematopoietic and lymphoid tissues in the CCLE samples. In agreement with the results for the bulk and single cell RNA-Seq datasets, chimera FUS390 was found to be highly expressed and chimera M19695 was determined to be lowly expressed in the K-562 cell line from the CCLE. The BCR-ABL1 chimera KT696168 was found in a total of eight cell lines of hematopoietic and lymphoid tissues. Chimeras FUS391 and FUS390 were seen in five cell lines, M19695 was noted in two cell lines, and FUS2195 and MH401088 were detected in only a single cell line.

We considered 2066 chimeras found in normal tissues of EBI as population reference chimeras. This population chimeras can be used to filter out chimeras that are frequently expressed in normal human tissues and do not associate with complex diseases, such as cancer. Next, 5956 cancer-specific chimeras found in CCLE cancer cell lines were used for downstream biological analysis. First, a list of unique fusion genes for each cancer was prepared to classify them as a driver gene, oncogene, or tumor suppressor gene. This list was uploaded to the CancerMine (47) database for annotation. Such annotation provides biological insight in terms of fusion genes and their classification as driver, oncogenic, or tumor suppressor genes for each cancer (S5 File). Moreover, 647 unique fusion genes from all cancers were uploaded into WebGestalt (48) for gene ontology analysis. The enrichment of these genes shows their significant association with pathways in cancers, such as the Ras signaling pathway, focal adhesion, and proteoglycans in cancer (FDR ≤ 0.05) (Supplementary Figure S6).

**DISCUSSION**

Access to an accurate and rapid chimera detection method is important both in cancer and complex disease research and for the precision medicine pipeline. With the advent of NGS technologies, global identification of chimeras became feasible. However, despite the development of over 20 computational methods designed to identify chimeras from high-throughput sequencing data, prediction sensitivity, specificity, accurate quantification of junction reads, execution time and hardware requirements remain challenging. Therefore, we developed ChiTaH, a fast ‘reference-based’ approach for the discovery of known human chimeras with superior accuracy. Currently, ChiTaH contains total of 43 466 non-redundant unique human chimeras combining exons of their parental genes, and thus with the potential of being translated into chimeric proteins.

In all scenarios tested, ChiTaH out-performed the current best methods for identifying chimeras. Specifically, in simulated datasets, ChiTaH out-performed these other methods in terms of sensitivity, specificity, estimating junction reads, execution time, and RAM required to complete the analysis (Figures 3 and 4). The same conclusion was drawn in the case of a true sequencing dataset from the K-562 cell line. The sole limitation in testing ChiTaH is that this tool cannot identify chimeras that are not available in its reference database of known human chimeras. Despite this limitation, the pipeline presented here is entirely customizable, such that user can easily incorporate known chimeras in which they are interested. Moreover, we have been updating our database ChiTaRS on a daily basis since 2012, resulting in addition of thousands of new known human chimeras annually (39, 49-51).

Finally, a unique future of ChiTaH reported here is the ability to identify variable junctions of the same chimera. Of all five tested methods, only ChiTaH discovered two distinct BCR-ABL1 chimeras in the K-562 cell line bulk RNA-Seq dataset, as well as in the single-cell RNA-Seq data. ChiTaH also accurately determined that BCR-ABL1 chimera FUS390 was significantly highly expressed, while chimera M19695 was poorly expressed in the bulk RNA-Seq dataset from K-562 cells. However, in the case of the single cell RNA-Seq dataset from K-562 cells, no significant differences between the expression of FUS390 and M19695 were noted, likely because chimera FUS390 was found to be expressed at only low levels in the majority of K-
Moreover, hundreds of RNA-Seq samples from EBI and CCLE were analyzed in just one month, showing that ChiTaH is well suited to meet the demands of large-scale tumor or population sample screening. In summary, the unique features of ChiTaH that allow it to utilize non-redundant junction sequence data for the identification of known chimeras also allow for the use of DNA-Seq or RNA-Seq sequencing datasets from bulk or single cell datasets. Finally, ChiTaH was also found to be efficient for the identification of sense-antisense (SAS) chimeras, given its unique reference-based approach (52).

CONCLUSIONS

In this report, we introduced ChiTaH, a fast and accurate method for the discovery of known human chimeras or fusion genes using DNA-Seq or RNA-Seq data generated using NGS technologies. ChiTaH, a user-friendly pipeline that will enable other research groups to make discoveries with ease of installation, is publicly available at GitHub (https://github.com/Rajesh-Detroja/ChiTaH).

DATA AVAILABILITY

Supplementary Data are available at NAR online. Programming codes, data and dependencies are available at GitHub (https://github.com/Rajesh-Detroja/ChiTaH).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Dr Sumit Mukherjee, Bar-Ilan University, Israel, for his suggestions on comparative analysis of chimera detection methods.

FUNDING

M.F-M. is supported by the Israel Innovation Authority [Kamin grant 66824, 2019–2021]. Conflict of interest statement. None declared.

REFERENCES
