GigaScience

Clinker: visualising fusion genes detected in RNA-seq data

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00019R2
Full Title:	Clinker: visualising fusion genes detected in RNA-seq data
Article Type:	Technical Note
Funding Information:	National Health and Medical Research Council (GNT1126157) — Prof. Alicia Oshlack National Health and Medical Research Council (APP1140626) — Prof. Alicia Oshlack National Health and Medical Research Council (GNT1145912) — Dr Ian Majewski Cancer Council Victoria (124178) — Dr Ian Majewski
Abstract:	Genomic profiling efforts have revealed a rich diversity of oncogenic fusion genes. While there are many methods for identifying fusion genes from RNA-seq data, visualising these transcripts and their supporting reads remains challenging. Clinker is a bioinformatics tool written in Python, R and Bpipe, that leverages the superTranscript method to visualise fusion genes 1. We demonstrate the use of Clinker to obtain interpretable visualisations of the RNA-seq data that lead to fusion calls. In addition, we use Clinker to explore multiple fusion transcripts with novel breakpoints within the P2RY8-CRLF2 fusion gene in B-cell Acute Lymphoblastic Leukaemia (B-ALL). Availability and Implementation: Clinker is freely available from Github https://github.com/Oshlack/Clinker under a MIT License.
Corresponding Author:	Alicia Oshlack AUSTRALIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Breon Schmidt
First Author Secondary Information:	
Order of Authors:	Breon Schmidt Nadia Davidson Anthony Hawkins Ray Bartolo Ian Majewski Paul Ekert Alicia Oshlack
Order of Authors Secondary Information:	
Response to Reviewers:	We have now made the formatting changes requested
Additional Information:	
Question	**Response**

Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Clinker: visualising fusion genes detected in RNA-seq data

Breon M Schmidt, Nadia M Davidson, Anthony DK Hawkins, Ray Bartolo, Ian J Majewski, Paul G Ekert, Alicia Oshlack

ABSTRACT

Summary: Genomic profiling efforts have revealed a rich diversity of oncogenic fusion genes. While there are many methods for identifying fusion genes from RNA-seq data, visualising these transcripts and their supporting reads remains challenging. Clinker is a bioinformatics tool written in Python, R and Bpipe, that leverages the superTranscript method to visualise fusion genes. We demonstrate the use of Clinker to obtain interpretable visualisations of the RNA-seq data that lead to fusion calls. In addition, we use Clinker to explore multiple fusion transcripts with novel breakpoints within the P2RY8-CRLF2 fusion gene in B-cell Acute Lymphoblastic Leukaemia (B-ALL).
Availability and Implementation: Clinker is freely available from Github https://github.com/Oshlack/Clinker under a MIT License.
Contact: alicia.oshlack@mcri.edu.au, ORCID: 0000-0001-9788-5690

Introduction

Genomic structural abnormalities, such as translocations between and within chromosomes, are common in cancer and can result in the fusion of two genes which then function as an oncogenic driver. The first example of this was the recurrent t(9;22) fusion in Chronic Myeloid Leukaemia, creating the *BCR-ABL1* oncogene [1]. This fusion gene results in a constitutively activated tyrosine kinase protein that can be effectively treated with small molecule inhibitors of ABL1, such as imatinib and dasatinib [2]. The application of next generation sequencing in cancer, primarily transcriptome sequencing (RNA-seq), has subsequently identified thousands of different fusion genes in many cancer types [3].

While there are many methods available for identifying fusion genes from RNA-seq data, there are few ways to visualise the fusion transcripts and the sequencing reads that support them. Simply aligning RNA-seq data to a reference genome or transcriptome does not allow clear visualisation of the translocation, or an appreciation of additional features such as splice variants. One approach for visualizing the translocation is to use the split screen view within IGV (Figure 2A). However, because RNA-seq read coverage is sparse in the genome, visualisation is hampered by the presence of introns. Other strategies that address this problem involve using predicted breakpoints to create the fusion transcript sequence, which can be used as a reference for read alignment [4,5]. This approach demonstrates coverage across the fusion breakpoints but other information about the structure and expression of the fusion transcripts, such as its expression relative to non-fused transcripts, can be lost. In addition, intronic sequence can be "shrunk" to give a more informative view of coverage[6].

Here we provide an alternative tool, Clinker, for visualizing RNA-seq data of fusion genes that enables a greater understanding of transcript coverage and splicing isoforms.

Clinker, utilises superTranscripts, a new type of transcriptome reference we previously developed that contains only the transcribed sequence of a gene, without introns, providing a highly compact reference for analysis and visualisation of RNA-Seq [7]. Clinker uses the human superTranscript references and creates fusion-superTranscripts by combining the two genes involved in a fusion event.

We have applied Clinker to a set of six B-cell ALLs that all report the *P2RY8-CRLF2* fusion to demonstrate several fusion isoforms. Clinker is a tool that provides direct visualisation of fusion genes and allows further appreciation of their complexity, such as alternative fusion isoforms.

Materials & Methods

Reference & Annotation Generation

The Clinker pipeline takes output from any fusion calling software, providing that it includes the hg19 or hg38 genomic coordinates of fusion gene breakpoints. These breakpoints are used to identify the two genes involved in the fusion and assigns them consistent gene symbols. This method is preferred even when gene symbols are provided by the fusion caller, due to the large variability in gene naming conventions. Once the two genes are identified, their sequences are retrieved from Clinker's human superTranscript reference and concatenated to form a single fusion-superTranscript. An important feature of the fusion-superTranscript reference is that it includes the full sequence of both genes orientated in transcriptional direction. Thus, reads aligned to regions of the genes not involved in the fusion are also visualised, providing additional information about expression of these regions and the domains they encode. This is repeated for all fusion genes that have been identified in the sample. This results in a sample specific Clinker reference containing the fusion-superTranscripts, as well as the superTranscripts from all normal genes. We found it was important to map competitively to the non-fused genes in the reference to avoid spurious read alignments. Transcript, protein domain and gene boundary annotation files are also created using the Gencode24 hg38 reference [8] and the Pfam protein database [9] to provide additional information for the visualisation.

Alignment to the new reference

Clinker maps the sequencing reads to the newly generated reference using the STAR aligner (STAR, RRID:SCR_015899)[10]. The aligner must be splice aware as reads spanning the fusion breakpoint are identified as splice sites. The alignment stage of Clinker often yields greater read support for fusion genes than fusion callers. For example, in one sample JAFFA detected the *P2RY8-CRLF2* fusion with support from 53 spanning reads whereas Clinker, through STAR, reported 290 spanning reads (Table S1). As Clinker is given prior knowledge that a fusion exists between two genes, the fusion-superTranscript can be mapped to with less stringency, leading to the increase in successfully mapped reads across the breakpoint. After the alignment step, the mapped reads and fusion genes can be viewed with a genome viewer, such as IGV, by loading the Clinker reference FASTA, mapped reads, and the customised transcript, protein domain and gene boundary annotation tracks. IGV natively

displays the fusion breakpoints and splice junctions through the splice junction track or sashimi plot (Figure 2B).

Filtering, normalisation and figure creation

Once reads are aligned, filtering and normalisation steps are undertaken. Split reads with a small number of flanking bases on one side can be produced by incorrect split-read alignment. To account for this split reads with less than 5 base pairs of flanking sequence are immediately filtered out using both Samtools (SAMTOOLS , RRID:SCR_002105)[11] and custom AWK scripting (see Supplementary Figure 2 for an example). Coverage is normalised to reads per million (RPM) using STAR's inbuilt normalisation function to allow comparison between samples.

A series of figures, one for each of the identified fusion genes, are then created using the R package, GViz [12]. These figures contain multiple tracks including coverage, gene boundaries, protein domains and the transcripts/exons involved in the fusion gene. A sashimi plot is also included in the figure to indicate the number of split reads that support the fusion, with three reads being set as a minimum threshold to further filter out spurious splicing events. The order, colour or presence of the tracks can be customised via the command line parameters of Clinker.

Software requirements

Clinker can be run both manually and through Bpipe (Bpipe, RRID:SCR_003471)[13], a tool for running bioinformatics pipelines. The core dependencies for Clinker are STAR [10], Samtools [11] and Gviz [12]. Runtime was approximately 1 hour with 8 processors and 40 GB of memory allocated for a single publication quality figure and a further 1 minute for each additional figure. This test was conducted on a sample with approximately 130 million reads and 2007 fusion genes reported by JAFFA.

***P2RY8-CRLF2* cloning and expression**

We have applied Clinker to a set of six B-cell ALL patient samples for which the *P2RY8-CRLF2* fusion gene was detected (Figure S1). We also found when using the JAFFA fusion caller several non-canonical fusion isoforms were reported. RNA-seq data for these sample can be found at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113504.

Fusion detection by RNA-seq was confirmed by PCR using gene specific primers for *P2YR8* (5'-CAAGGTTGCTGGACAGATGGAA-3') and *CRLF2* (5'-AATAGAGAATGTCGTCTCGCTGC-3'). Primers were designed to amplify products spanning the exons at the breakpoints of *P2RY8* and *CRLF2* in the mRNA transcripts detected by JAFFA (Figure S3). The alternate and frameshift fusions were cloned using primers to target the start of P2RY8 (5- CCCTGCACATGAGTGTTCAGAC-3') and the end of *CRLF2* (5'- TCACAACGCCACGTAGGAG -3'), while the canonical fusion was amplified using a different *P2RY8* forward primer (5'- GCGGCCGCCTTTGCAAGGTTGC-3') (Figure S4). PCR products were cloned into P-GEM-T easy vector (Promega), Sanger sequenced and then subcloned into a retroviral pMSCV-GFP retroviral expression vector. Retrovirus was produced as previously described [14], and transduced into IL3-dependant BaF3

cells. *CRLF2* was detected in the BaF3 cells using the Anti-Human TSLP receptor antibody (eBioscience) and the BD Cytofix/Cytoperm (BD Biosciences), according the manufacturer's instructions (Figure S5). FACs analysis was performed on an LSRII flow cytometer (BD Biosciences).

Results

The Clinker pipeline

Clinker is an analysis pipeline that takes in fusion calls and raw RNA-seq reads and outputs a custom reference, mapped read data and image files to visualise and assess fusion transcripts. The steps in this pipeline are outlined in Figure 1 and described in detail in the Materials and Methods.

Briefly, before running Clinker, fusions are detected using one of the many specialised fusion gene callers, such as JAFFA [15], STAR-fusion [16] or Pizzly [17]. Clinker proceeds by first concatenating the full-length superTranscripts of the two genes involved in the fusion for each event called in the sample. These fused superTranscripts are then added to a custom, sample specific superTranscriptome reference. Next, the reads are mapped back to the new reference using the STAR splice aware aligner [10]. A fusion can then be observed as splicing between the two concatenated genes. Finally, figures are generated that present the resulting splice junctions, coverage, protein domains and transcript annotation, for both the fusion and non-fused superTranscripts. Clinker outputs file formats that are compatible with IGV [18], as well as publication quality images created with Gviz [12].

Clinker visualisations of reads, transcripts and protein domains

Most fusion calling algorithms use short-read RNA-seq data to report genes involved in potential fusion events as well as the number of reads detected that support these events. Figure 2 demonstrates the visualization of an *KMT2A-MLLT3* fusion gene that was detected in a B-cell ALL using JAFFA [15] with and without using Clinker. Visualizing this fusion using IGV without Clinker is done using a split screen display of the regions of the genome spanning the fusion breakpoints (Figure 2A). While read pairs that span across the fusion breakpoints are viewable (green reads), the transcript context is difficult to discern. In contrast, using the Clinker superTranscript reference and outputs allows a neater and more informative visualization in IGV which can also display sashimi plots for the fusion support (Figure 2B). Finally Clinker also outputs a PDF image of the fusion that can be customised (Figure 2C).

Identification of novel fusion isoforms in P2RY8-CRLF2

In order to demonstrate the utility of Clinker to provide visualization and insight into fusion genes we applied Clinker to six B-Cell Acute Lymphoblastic Leukaemia (B-ALL) samples that carried the *P2RY8-CRLF2* fusion. This fusion gene is reported to be present in ~7% of B-ALL cases and results in the overexpression of *CRLF2* [19]. The canonical fusion

joins the first non-coding (UTR) exon of *P2RY8* to the start of the coding region of *CRLF2* [19]. Commonly, this fusion arises as a result of an interstitial deletion in the Par1 region of chrX or chrY [19]. Interestingly, JAFFA called multiple breakpoints in *CRLF2* for this fusion gene in each of the sequenced B-ALL samples, suggesting different isoforms of this fusion. JAFFA identified the canonical break point in all samples. In addition, each of the six samples also expressed an isoform of *P2RY8-CRLF2* which joined the first exon of *P2RY8* to the 5' UTR of *CRLF2*, and resulted in an in-frame transcript. This alternate fusion also featured the typical GT/AG donor/acceptor motif which exists at the majority of splice junctions [20]. The presence of the alternate fusion isoforms in the samples was confirmed using RT-PCR and Sanger sequencing.

We used Clinker to visualize the *P2RY8-CRLF2* fusions in all six samples (Figure 3). Clinker detected the canonical breakpoint with the highest coverage in all samples (blue lines) and the novel 5' splice site in all samples at much lower levels (red lines). Interestingly, all six samples had additional splice sites between exon 1 and 2 of *CRLF2* (green lines) and one sample had a fourth isoform with a splicing breakpoint between exon 5 and 6 of *CRLF2* (yellow lines). However these additional transcripts are not predicted to be in-frame.

To determine if the alternative, low abundance, in-frame transcript could drive *CRLF2* over expression, and so potentially contribute to the biology of ALL driven by *P2RY8-CRLF2* fusions we cloned the canonical and alternative version of the *P2RY8-CRLF2* fusion, as well as a frameshift version to act as a negative control, into retroviral vectors. The erythroleukaemia cell line (BaF3 cells) were then transduced with these retroviruses to produce cell lines constitutively expressing the in-frame fusions or the negative (frameshift) *P2RY8-CRLF2* control. We measured CRLF2 expression using an anti-human CRLF2 antibody and flow cytometry (Supplementary Figure S5). The data show that both the canonical fusion and the alternate in-frame transcript can drive CRLF2 overexpression in BaF3 cells, but that the shorter frame shift transcript does not. These data suggest that alternate transcript isoforms can contribute to the overexpression of CRLF2 in B-ALL.

Discussion

Here we present Clinker, a visualisation tool for exploring and plotting fusion genes discovered in RNA-seq data. Clinker uses the idea of superTranscripts to build a reference for identified fusions allowing the raw reads involved in the discovery of fusion genes to be viewed and inspected in IGV. Mapping reads back to this generated fusion gene reference generally results in greater read support for true fusion events. In addition Clinker annotates transcripts and protein domains providing far greater insight into the expression levels and structures of the transcripts that make up the fusion gene. Publication quality figures can be easily generated and refined using R functions. Appling Clinker to real data demonstrated that alternative splicing could be detected within a single fusion gene. Our examination of the *P2RY8-CRLF2* fusions indicates that these alternate isoforms exist in the primary samples, and may have biological relevance, as it appears capable of encoding a functional CRLF2 protein.

Availability of supporting source code and requirements

- Project name: Clinker
- Project home page: https://github.com/Oshlack/Clinker/
- Operating systems: 64 bit Linux or Mac OS X
- Programming language: Python, R, Bash
- Other requirements: STAR, Samtools and Gviz.
- License: MIT
- RRID: Clinker, RRID:SCR_016140

Availability of supporting data

Snapshots of the code are available from the *GigaScience* GigaDB repository [21].

Competing interests

The authors declare that they have no competing interests

Acknowledgements

This work was supported by grants from the Australian National Health and Medical Research Council (NHMRC) (Project Grants to AO 1140626, IJM 1145912; CDF (to AO) 1126157, Independent Research Institutes Infrastructure Support Scheme grant 9000220), the Cancer Council Victoria (grant-in-aid to IJM 1124178), a Victorian State Government Operational Infrastructure Support (OIS) grant; a Victorian Cancer Agency fellowship (to IJM) and a Felton Bequest to IJM.

References

1. Shtivelman, E., Lifshitz, B., Gale, R. P. & Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* **315,** 550 (1985).

2. Quintás-Cardama, A., Kantarjian, H. & Cortes, J. Imatinib and beyond—exploring the full potential of targeted therapy for CML. *Nat. Rev. Clin. Oncol.* **6,** 535–543 (2009).

3. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat. Rev. Cancer* **15,** 371–381 (2015).

4. Beccuti, M. *et al.* Chimera: a Bioconductor package for secondary analysis of fusion products. *Bioinformatics* **30,** 3556–7 (2014).

5. Lågstad, S. Chimeraviz: Visualization tools for gene fusions. (2017).

6. Haas, B. & Tickle, T. Fusion Inspector. *Github* (2017). Available at: https://github.com/FusionInspector/FusionInspector.

7. Davidson, N. M., Hawkins, A. D. & Oshlack, A. SuperTranscript: a data driven reference for analysis and visualisation of transcriptomes. *bioRxiv* (2017).

8. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–74 (2012).

9. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279-85 (2016).

10. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

12. Hahne, F. & Ivanek, R. in 335–351 (2016). doi:10.1007/978-1-4939-3578-9_16

13. Sadedin, S. P., Pope, B. & Oshlack, A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28,** 1525–6 (2012).

14. Narayan, N. *et al.* Functionally distinct roles for different miR-155 expression levels through contrasting effects on gene expression, in acute myeloid leukaemia. *Leukemia* **31,** 808–820 (2017).

15. Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* **7,** 43 (2015).

16. Haas, B. *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv* (2017).

17. Melsted, P. Pizzly. *Github* (2017). Available at: https://github.com/pmelsted.

18. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).

19. Mullighan, C. G. *et al.* Rearrangement of CRLF2 in B-progenitor– and Down syndrome–associated acute lymphoblastic leukemia. *Nat. Genet.* **41,** 1243–1246 (2009).

20. Lai, J. *et al.* Fusion transcript loci share many genomic features with non-fusion loci. *BMC Genomics* **16,** 1021 (2015).

21. Schmidt, B; Davidson, N; Hawkins, A; Bartolo, R; Majewski, I; Ekert, P; Oshlack, A (2018): Supporting data for "Clinker: visualising fusion genes detected in RNA-seq data" GigaScience Database. http://dx.doi.org/10.5524/100460

FIGURE CAPTIONS

Figure 1. A visual representation of the Clinker pipeline. Users can choose to stop at step two, inspect fusion genes of interest in IGV, and then commence figure production for a refined list of fusion genes. The Fusion Caller (pink) and IGV (light blue) steps are external to Clinker.

Figure 2. *KMT2A-MLLT3* fusion gene visualised in IGV after alignment to the human genome (A). The backgrounds of the IGV tracks are coloured to distinguish between the coverage (purple), aligned reads (white) and annotation (blue), with green reads indicating that its partner is on a different chromosome. Such alignments may support the existence of a fusion. (B) Clinker output of the *KMT2A-MLLT3* gene fusion, visualised in IGV and (C) the GViz visualisation. The tracks in the Clinker GViz visualisation are (top to bottom): a superTranscript scale axis, a read coverage track, a gene boundary track, a protein domain track, a transcript (with exons annotation) track and a sashimi plot that indicates the fusion breakpoints (dark purple). The breakpoints are also indicated by the vertical lines. In addition to the Clinker tracks, the IGV visualisation includes a read support track.

Figure 3. Visualisation of six samples containing the P2RY8-CRLF2 fusion. We combined the Clinker output (mapped reads, fusion superTranscript and annotation track) for the six samples using Gviz in R. From top to bottom: six coverage tracks with annotated breakpoints demonstrating read support, gene track, protein domains and gene transcripts. Each sample contains the canonical transcript (navy vertical line) as well as a novel upstream splicing occurring within the 5'UTR exon of CRLF2 (annotated with the red vertical line over the CRLF2 gene) along with two other transcripts that are not in frame. The read support for this Clinker output can be compared to that of JAFFA's in Supplementary Tables 1.

Figure 1

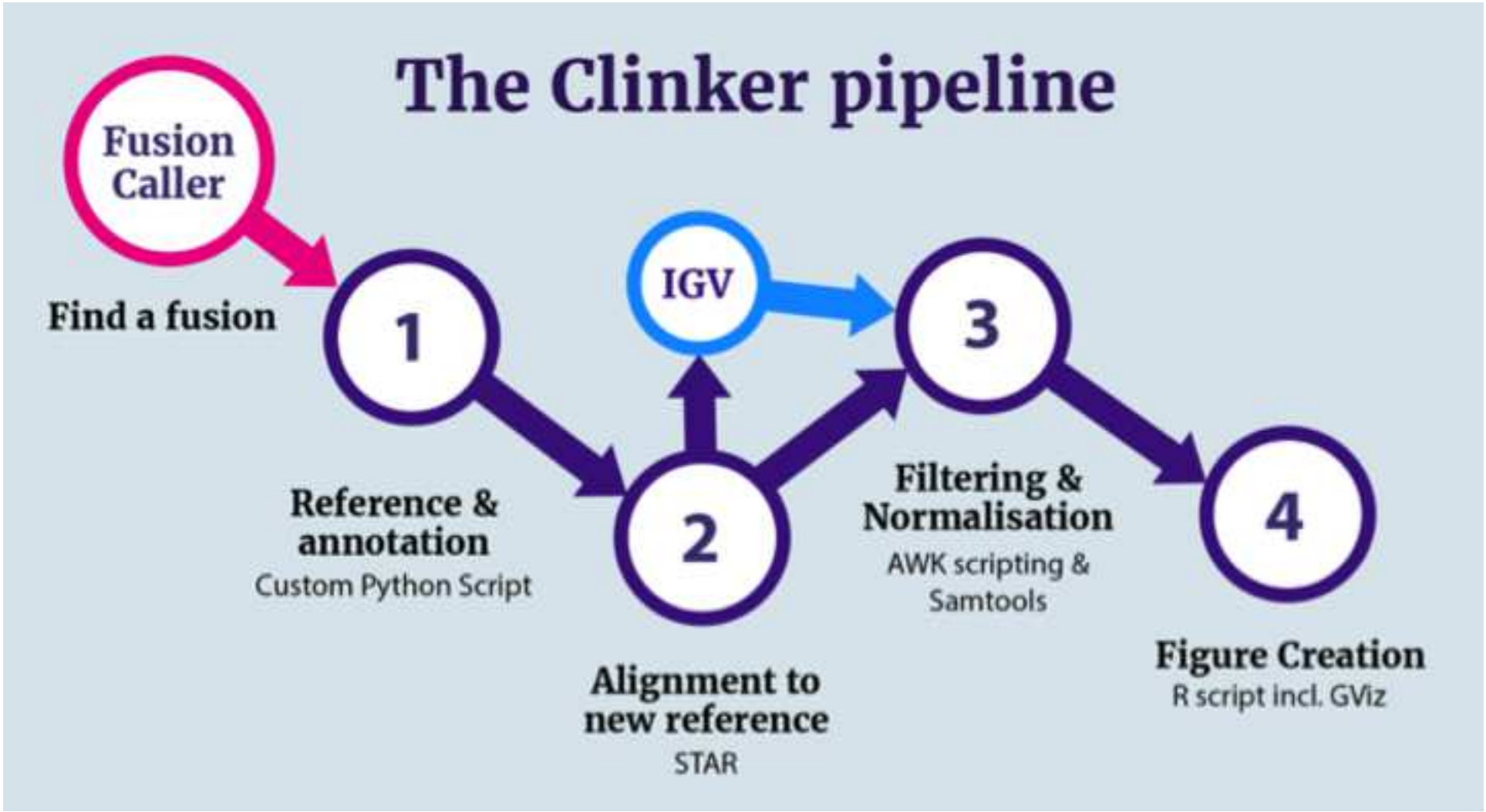
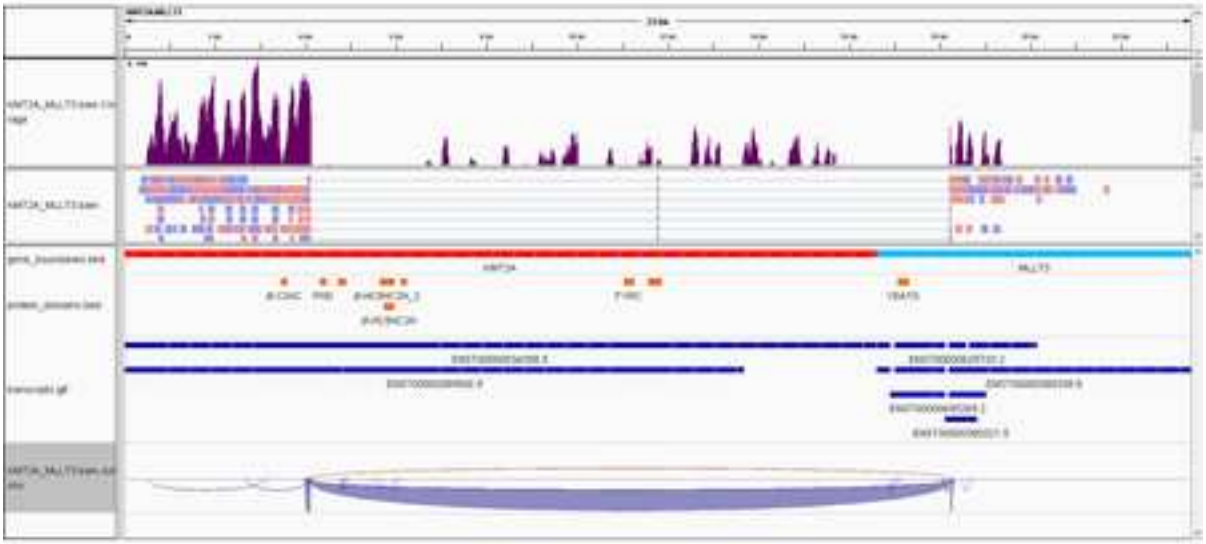


Figure 2

A) Alignment to Genome - IGV

B) Clinker Output - IGV



C) Clinker Output - Gviz

Figure 3

Click here to access/download;Figure;Fig3.png