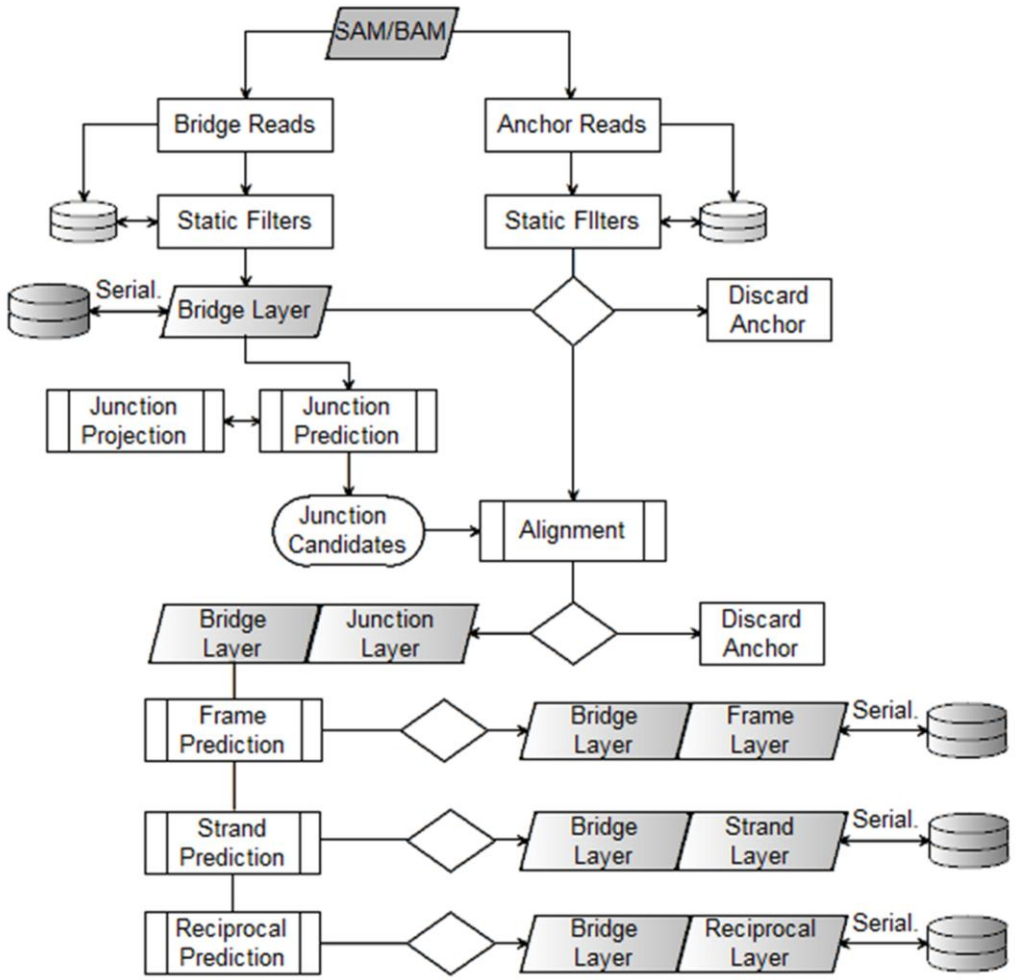
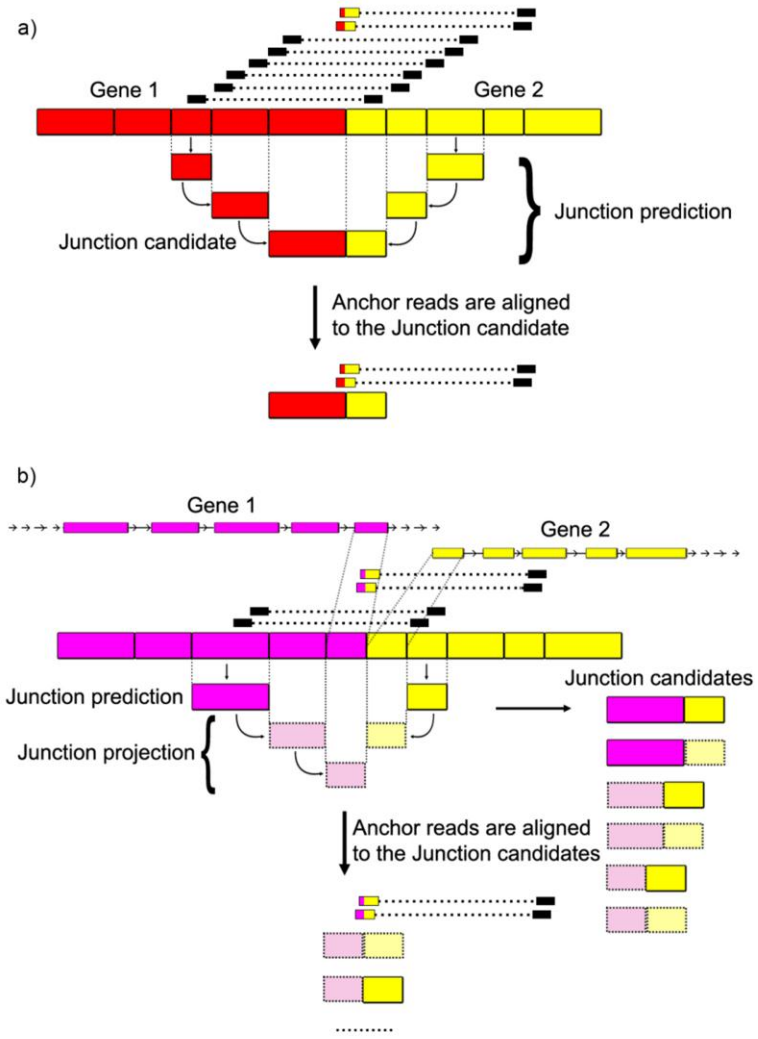
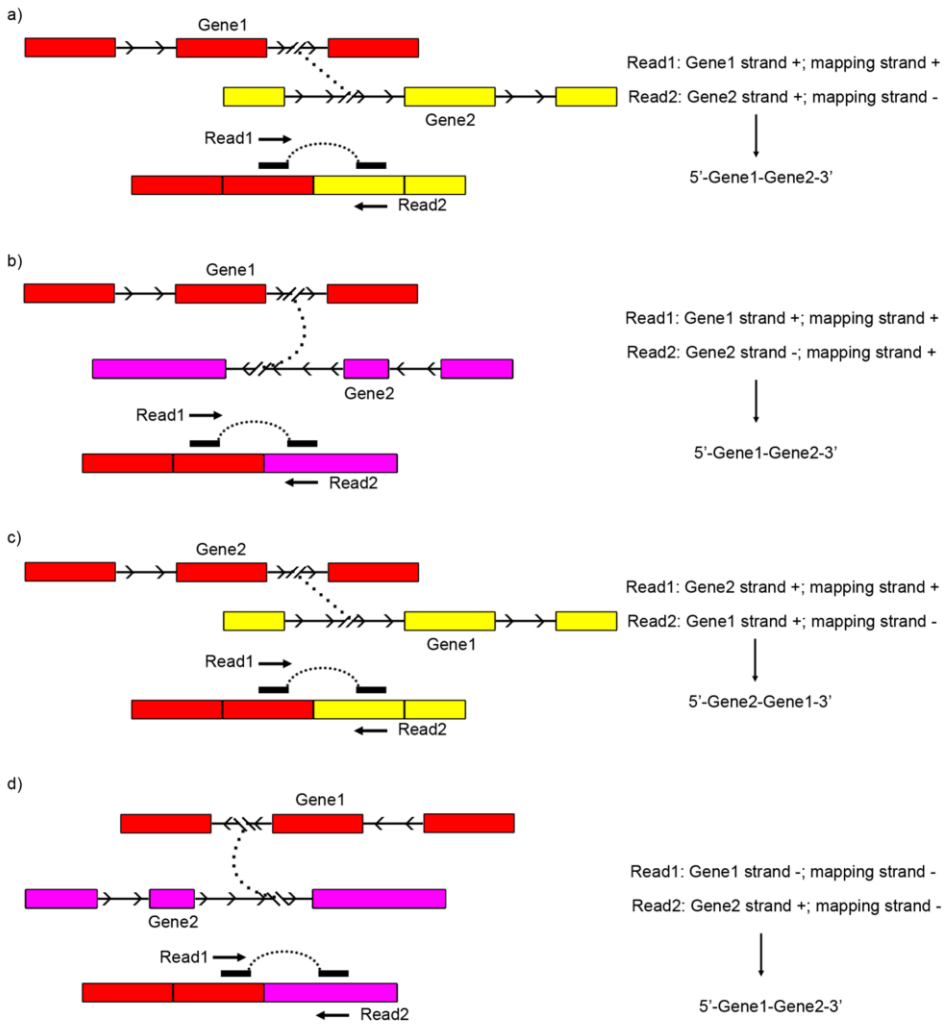


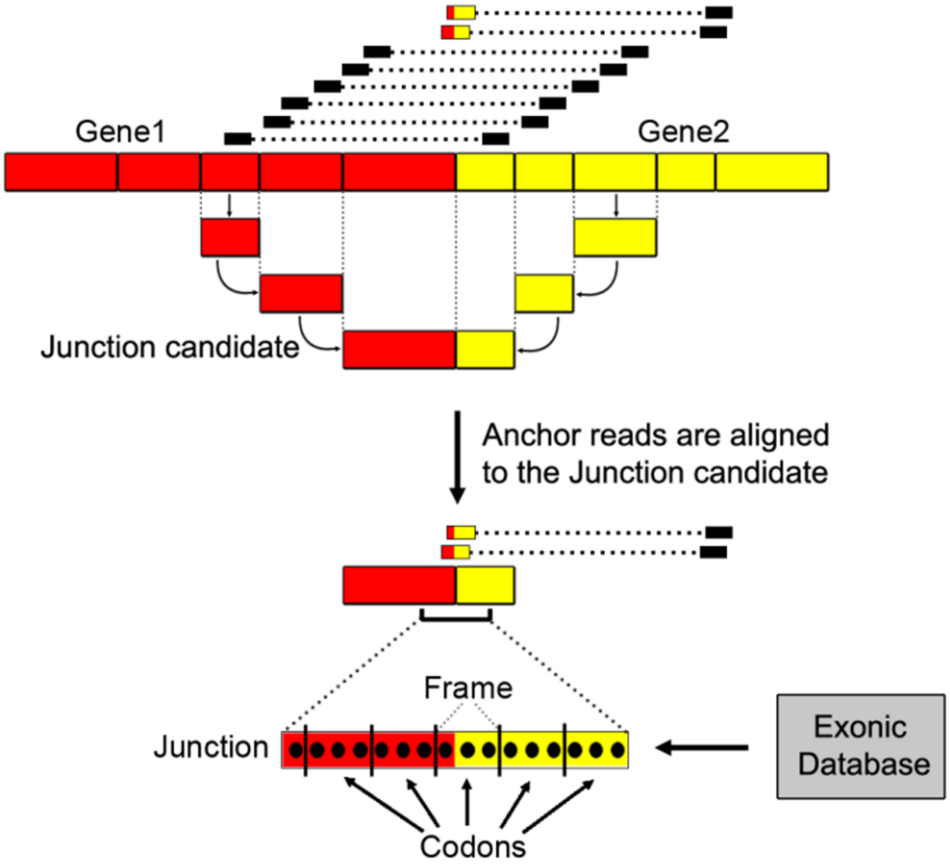
Suppl. Fig. 1: schematic flowchart of the FusionAnalyser algorithms. The ‘Serial’ label indicates a Serialization/Deserialization process. The grey cylinders indicate non-volatile storage systems.



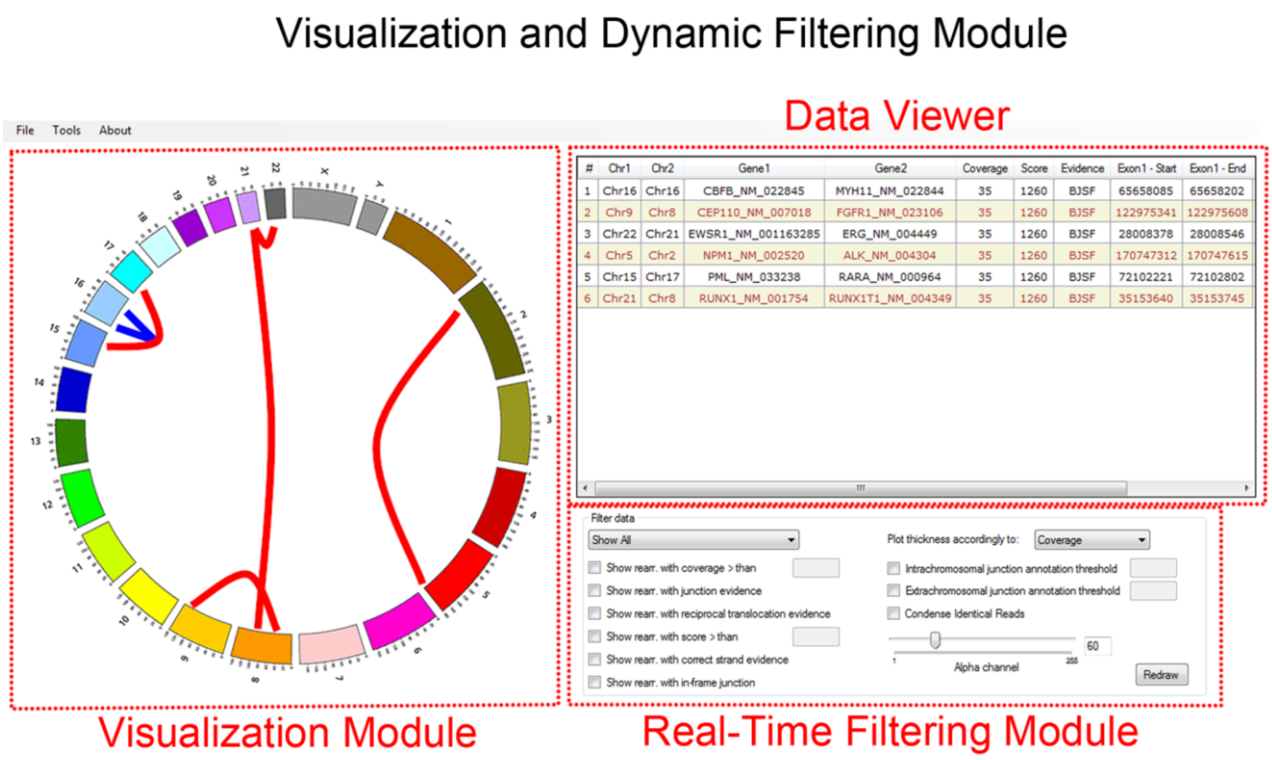
Suppl. Fig. 2: a) Outline of the Junction Prediction algorithm. The red and yellow boxes represent individual exons of a generic fusion. The black rectangles separated by a dashed line indicate the position of the Bridge reads. The black and yellow/red rectangles separated by a black dashed line indicate the position of the Anchor reads. The Junction candidate is generated by identifying all the exons of each partner being aligned to one or more Bridge reads. If one of the two partner genes is at the 5' of the fusion (Gene1), according to the Strand prediction algorithm, the Gene1-exon contributing to the Junction candidate will be the 3'-most exon among all those receiving the alignment of at least one Bridge read. If the partner gene is at the 3' of the fusion (Gene2), the Gene2-exon contributing to the Junction candidate will be the 5'-most exon among all those receiving the alignment of at least one Bridge read. b) Outline of the Junction Projection algorithm. The Junction Projection algorithm takes the Gene1 and Gene2 candidate breakpoint exons (violet and yellow boxes) identified by the Junction Prediction algorithm and the strand prediction as input and projects new fusion candidates (light violet and light yellow dotted boxes), in absence of Bridge reads mapping to them, according to the required depth of the projection. The candidate breakpoint fusions are generated combinatorially by combining all the Gene1-exon and Gene2-exon candidates. The candidate breakpoint fusions are then aligned against all the Anchor reads (black and violet/yellow rectangles separated by a black dashed line) mapping to Gene1 or Gene2.



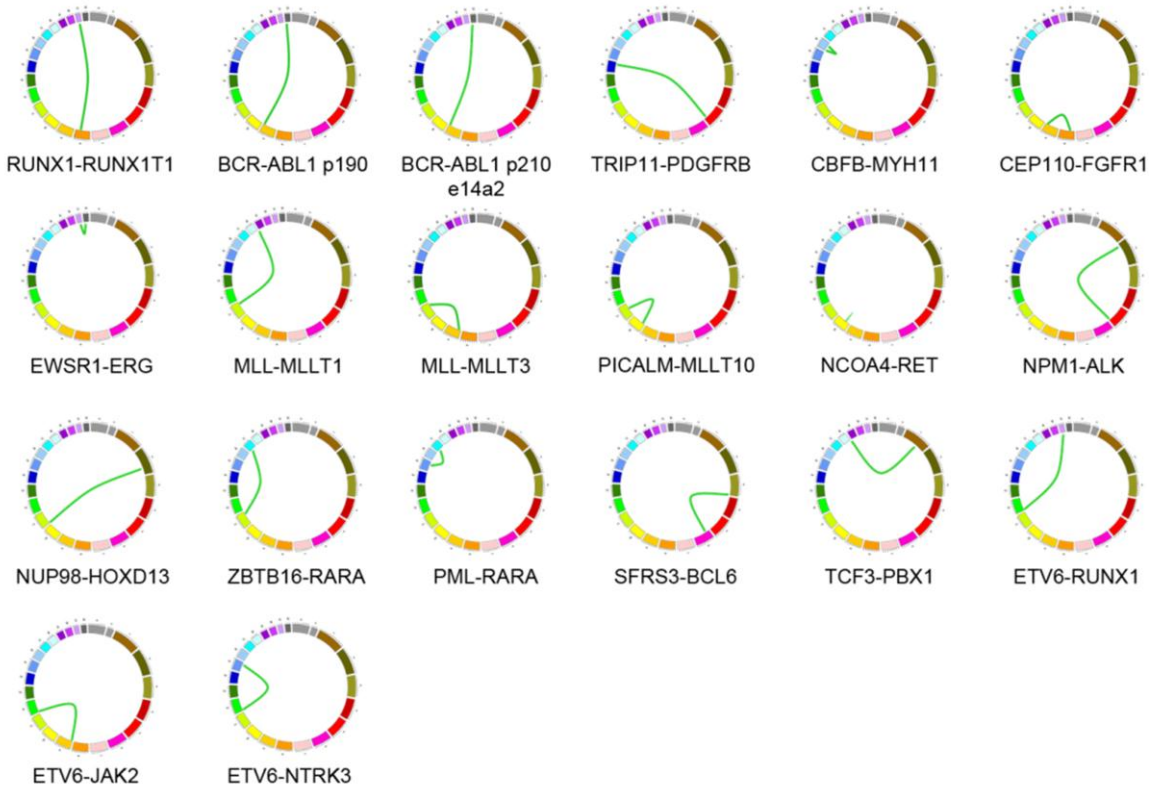
Suppl. Fig. 3: schematic modeling of four (a, b, c, d) Gene and Mapping Strand combinations and the resulting fusion output. Red, Yellow and Purple boxes indicate exons. The arrowed lines indicate the direction of the coding strand of each gene. The black rectangles separated by a dashed line indicate individual Bridge reads. The nearby black arrows indicate the direction of the sequencing. On the right, the gene and mapping strand of each model and the resulting fusions are summarized.



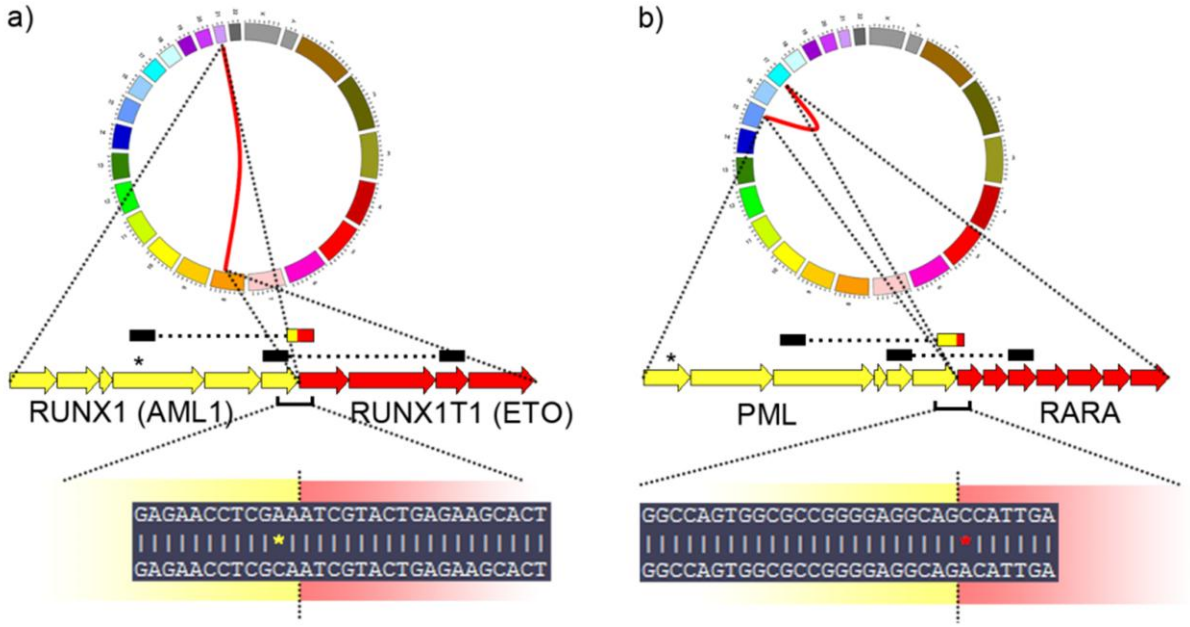
Suppl. Fig. 4: the process of frame annotation relies on the initial identification of one or more candidate junctions (upper panel) and on the subsequent definition of a valid breakpoint junction (middle panel), supported by Anchor reads. The two breakpoint exons are then analyzed in the context of their coding sequence through the interaction with an exonic database in order to retrieve their coding frame. The frame compatibility between the two exons is then calculated. Red and yellow boxes represent exons. The black rectangles separated by a dashed line indicate the position of the Bridge reads. The black and red/yellow rectangles separated by a dashed black line indicate the position of the Anchor reads mapping to the breakpoint. In the lower panel, the black circles represent single nucleotides; the vertical black lines separate individual codons.



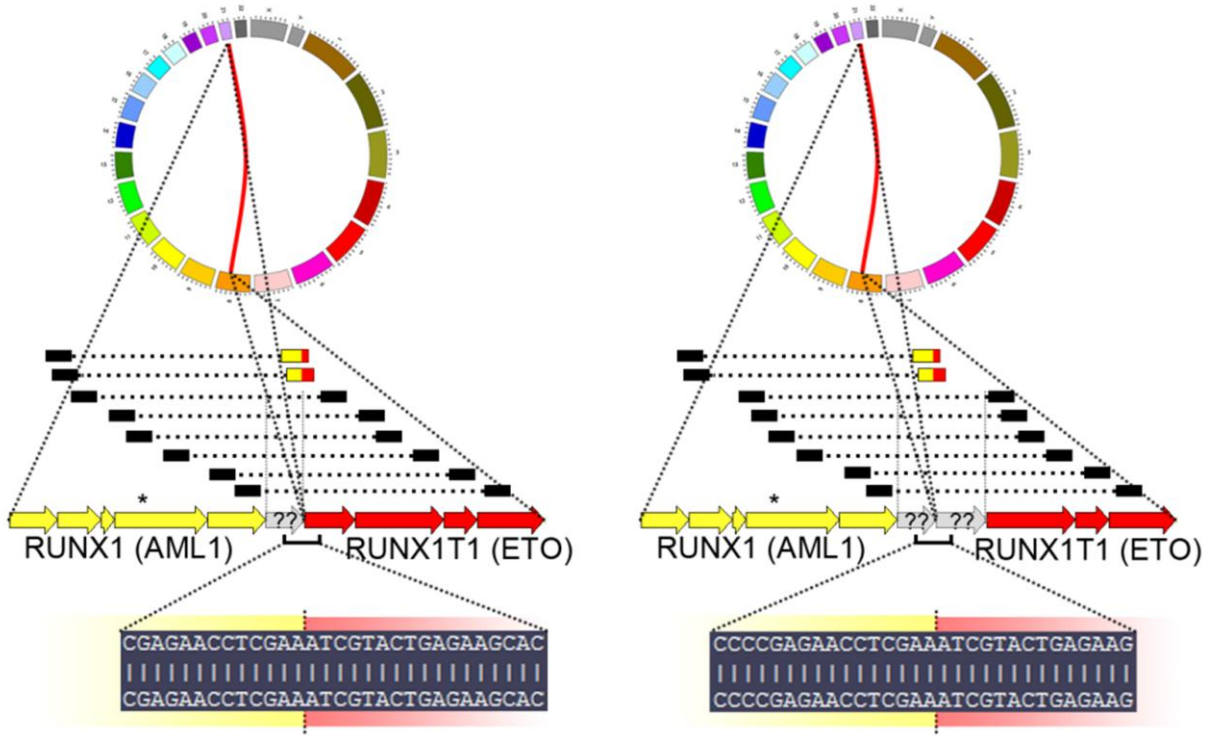
Suppl. Fig. 5: Visualization and Dynamic Filtering Module. The VIDYF module represents the standard graphical FusionAnalyser output module. It is composed by 3 main panels: 1) the Visualization module (left panel), where the graphical visualization of the filtered fusions take place. The visualization is user customizable: under the standard configuration, the thickness of each fusion line is proportional to the coverage of the reads pointing to that fusion. The alpha blending is inversely proportional to the normalized score of each fusion. 2) the Data Viewer (top panel), where all the relevant information regarding each fusion can be retrieved, such as chromosome localization of the two rearranged genes, gene names, coverage, overall score, type of evidence (Bridge, Junction, Strand, Frame, Reciprocal) and genomic coordinates of the two breakpoint exons. The Data Viewer allows multiple dynamic interactions: data selection is directly reflected in the Visualization module and further information concerning individual fusions, such as the sequence of the breakpoint junction and exons, the list of Bridge and Anchor reads supporting an event, the alignment and score of each Anchor read, can be retrieved through popup menus. 3) the Real-Time Filtering module (bottom right panel). Here, intra and extrachromosomal candidate fusions can be dynamically filtered to select all the biologically relevant driver events in line with the following set of parameters: read coverage, overall scoring threshold, presence of Junction, Strand, Frame and Reciprocal translocation evidence, junction alignment score and removal of read duplicates. The fusion data generated through the dynamic filtering process are shown in real-time in the Visualization module and in the Data Viewer.



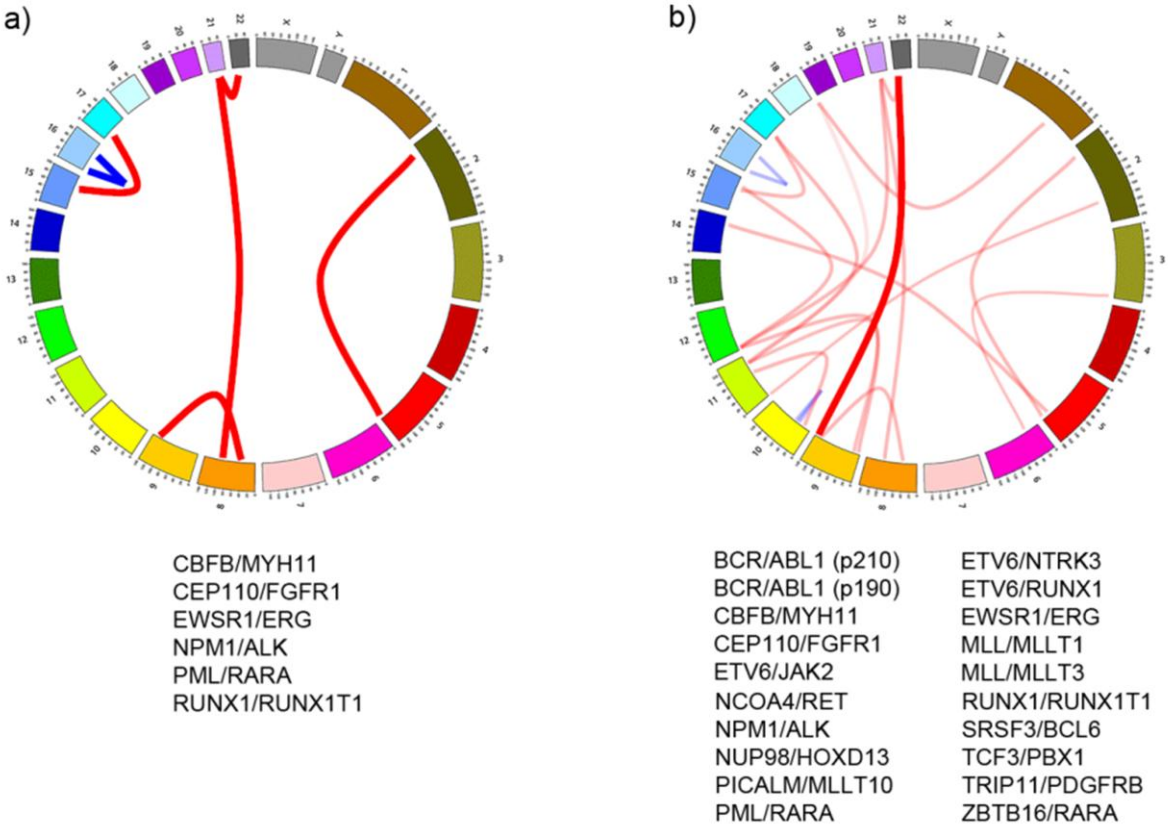
Suppl. Fig. 6: The graphical output of the FusionAnalyser analyses of 20 in silico rearrangement data is shown. The green lines connect the two chromosomes involved in each fusion.



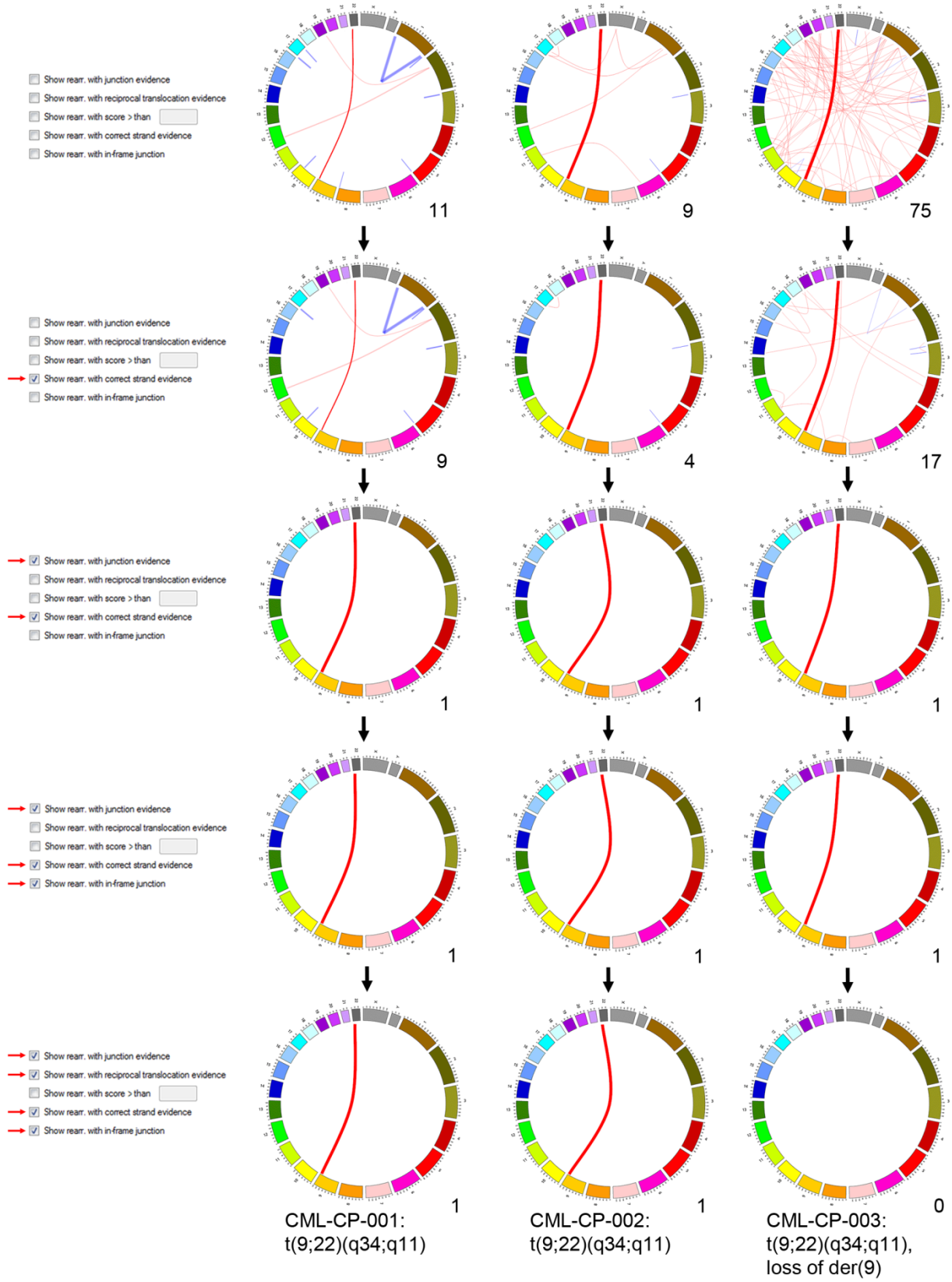
Suppl. Fig. 7: In the upper part of the panel, the output of a FusionAnalyser analysis generated by simulating low coverage fusion sequencing data and in presence of a single nucleotide mismatch in the context of the fusion region for RUNX1-RUNX1T1 (a) and PML-RARA (b) rearrangements is shown. The red line connects the two chromosomes involved in the t(8;21)(q22;q22) (a) and the t(15;17)(q24;q21) (b) fusions. In the middle panel, the schematic breakpoint regions are shown. The yellow arrows represent RUNX1 (a) or PML (b) exons; the red arrows represent RUNX1T1 (a) or RARA (b) exons. The asterisks indicate the beginning of the coding sequences. The two black rectangles separated by a dashed line indicate the position of the single Bridge read mapping to the RUNX1-RUNX1T1 (a) or PML-RARA (b) fusions. The black and yellow/red rectangles separated by a dashed black line indicate the position of the Anchor reads mapping to each breakpoint. In the lower panel, the breakpoint regions identified by FusionAnalyser are shown at nucleotide level. The upper sequence in each box represents the reference breakpoint sequence, generated by the Junction Prediction/Projection modules; the lower sequence represents part of the Anchor read successfully mapped to the breakpoint region despite the presence of a single nucleotide variant in the first (a, yellow asterisk) or second (b, red asterisk) breakpoint exons.



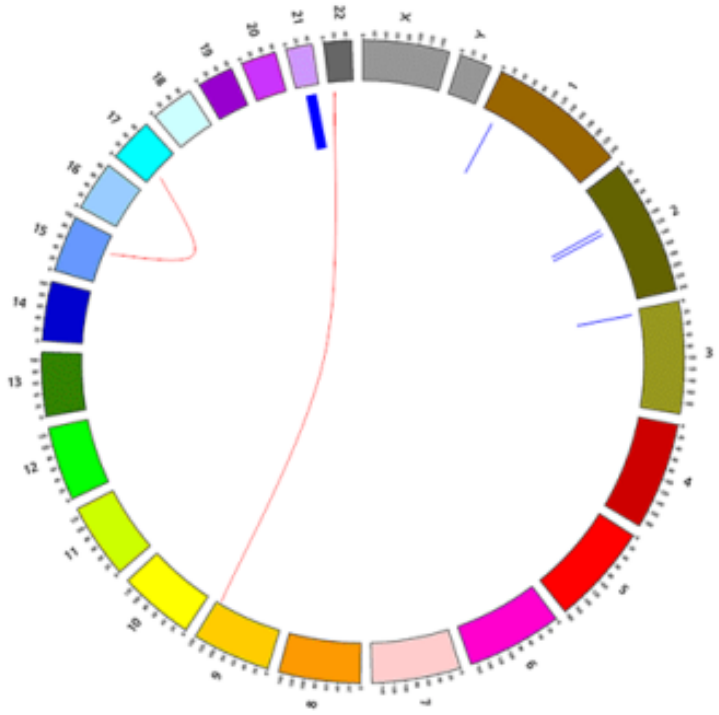
Suppl. Fig. 8: In the upper part of the panel, the output of a FusionAnalyser analysis generated by simulating the absence of Bridge reads mapping to the first (a) or to both the breakpoint exons (b) in the context of the fusion region for RUNX1-RUNX1T1 is shown. The red line connects the two chromosomes involved in the t(8;21)(q22;q22) fusion. In the middle panel, the schematic breakpoint regions are shown. The yellow arrows represent RUNX1 exons; the red arrows represent RUNX1T1 exons. The grey arrows represent RUNX1 or RUNX1T1 breakpoint exons not supported by the presence of Bridge reads. The asterisks indicate the beginning of the coding sequences. The two black rectangles separated by a dashed line indicate the position of the Bridge reads mapping to the RUNX1-RUNX1T1 fusion. The black and yellow/red rectangles separated by a dashed black line indicate the position of the Anchor reads mapping to each breakpoint. In the lower panel, the breakpoint regions identified by FusionAnalyser are shown at nucleotide level. The upper sequence in each box represents the reference breakpoint sequence, generated by the Junction Prediction/Projection modules; the lower sequence represents part of an Anchor read successfully mapped to the breakpoint region despite the absence of Bridge reads mapping to the breakpoint exons.



Suppl. Fig. 9: The graphical output of the FusionAnalyser analyses of 6 (a) or 20 (b) multiple in silico rearrangement data is shown. The red lines connect the two chromosomes involved in each fusion.



Suppl. Fig. 10: Analysis of transcriptome sequencing data of 3 CML patients. In the left part of the figure, the main FusionAnalyser real-time filters panel is shown. In the (a), (b) and (c) panels the individual fusion data of each CML patient are represented. From top to bottom, progressively more stringent criteria are applied: in the upper panel, all the BAL fusion candidates are shown. In the lower panels, the strand, junction, frame and reciprocal translocation filters are applied (red arrows). The numbers at the bottom-right of each circular diagram indicate the number of fusion candidates associated with the corresponding filter set.



Suppl. Fig. 11: Analysis of the combined CML002-AML002 transcriptome data. The two red lines highlight the presence of the PML-RARA and of the BCR-ABL1 fusions. The thick blue line on chromosome 21 points to the ETS2-ERG intrachromosomal fusion. The following (clockwise) thin blue lines point to four candidate readthrough events: TEKT2-ADPRHL2, VAMP8-VAMP5, ANKRD39-ANKRD23 and ARPC4-TTLL3, respectively.