# Alternative Splicing Events

**Dr. Giorgio Giurato, PhD**
ggiurato@unisa.it

Laboratory of Molecular Medicine and Genomics
Genomix4Life Srl
Department of Medicine, Surgery and Dentistry 'Schola Medica Salernitana'
University of Salerno

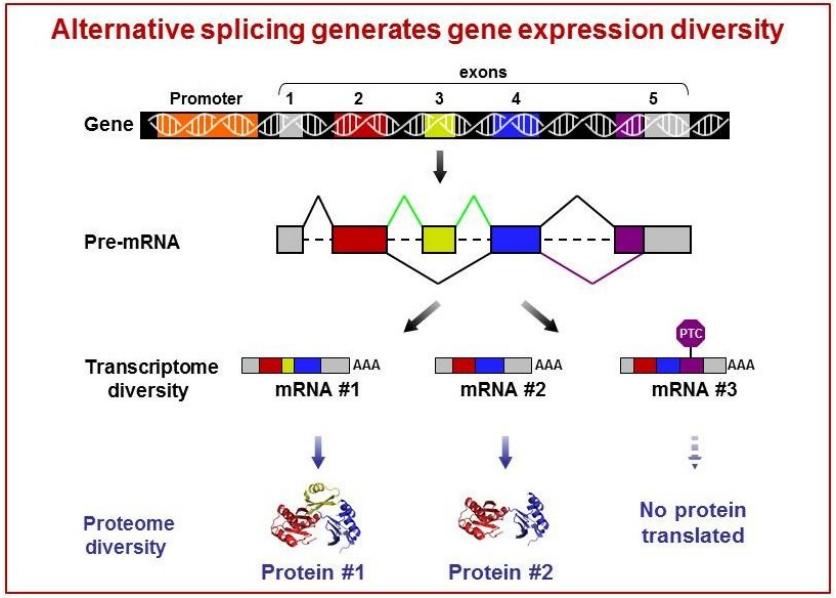
## Training Course on Best practise for RNA-Seq data analysis

Sept 27-29 2017 – Salerno

# What is Alternative Splicing?

- Alternative Splicing (AS) is a process by which exons or portions of exons or non-coding region within a pre-mRNA transcript are differentially joined or skipped, resulting in multiple protein isoforms.

- This mechanism increases the informational diversity and functional capability of a gene during post-transcriptional processing and provides an opportunity for gene regulation.

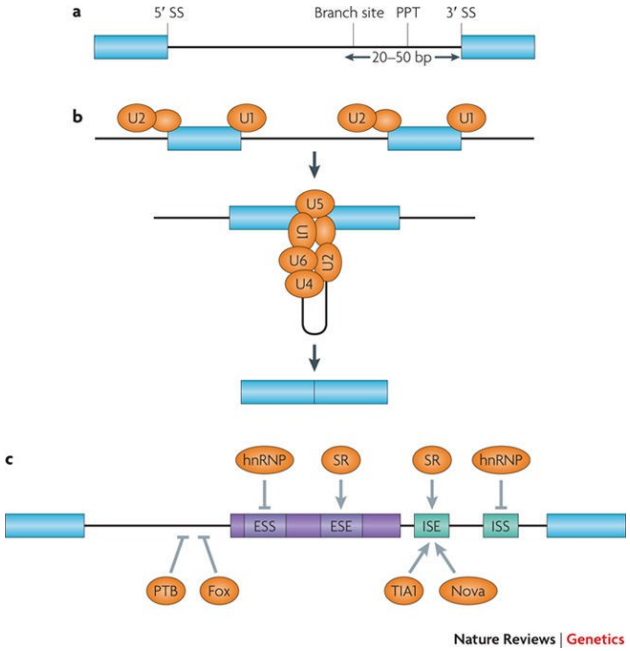


# The Splicing Machinery

- During AS, cis-acting regulatory elements in the mRNA sequence determine which exons are retained and which one ones are spliced out.

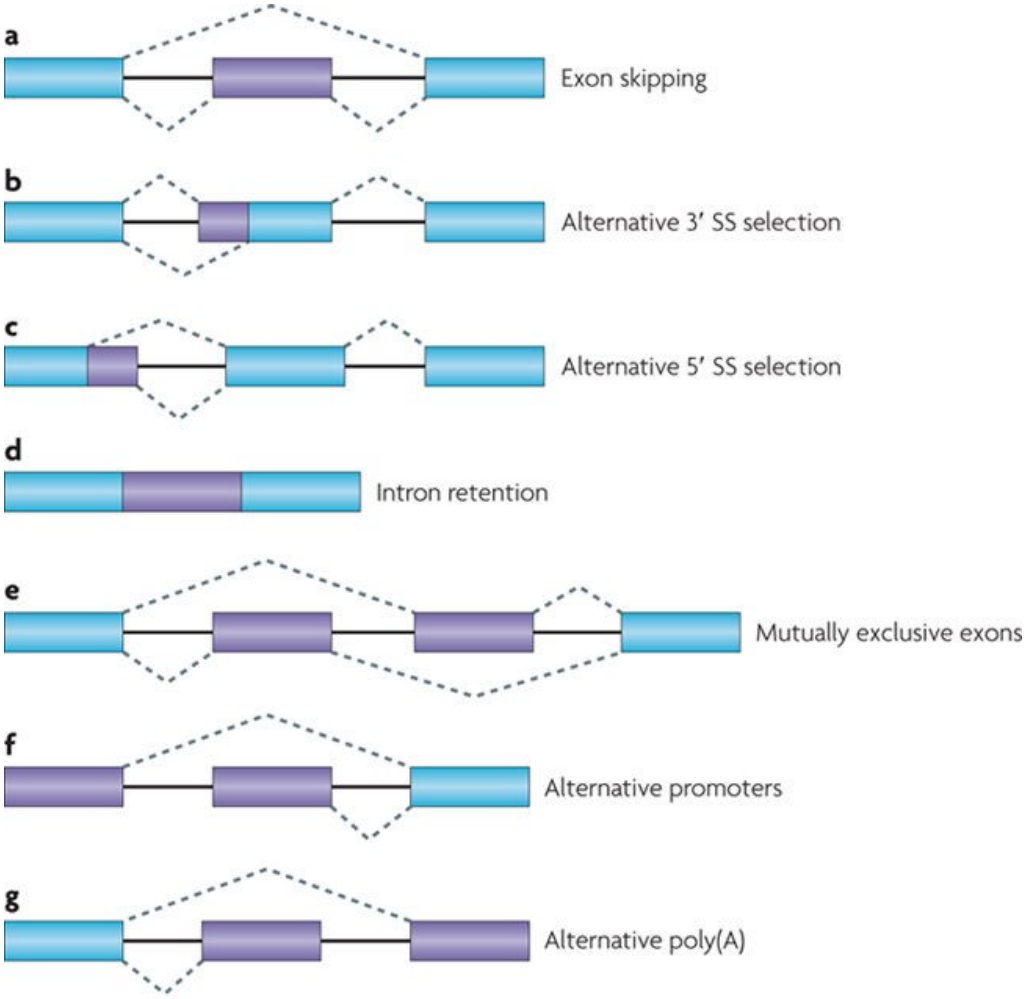
- These cis-acting regulatory elements alter splicing by binding different trans-acting protein factors.

- The final decision to include or splice an alternative exon is determined by combinatorial effect and cellular abundance.



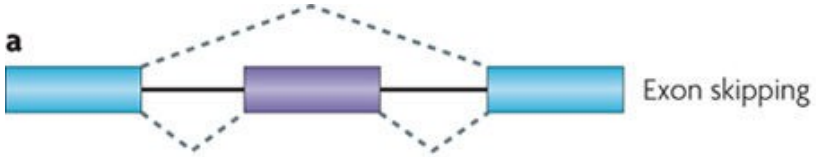
# Different types of Alternative Splicing

- There are several different types of Alternative Splicing events, which can be classified into four main subgroups.



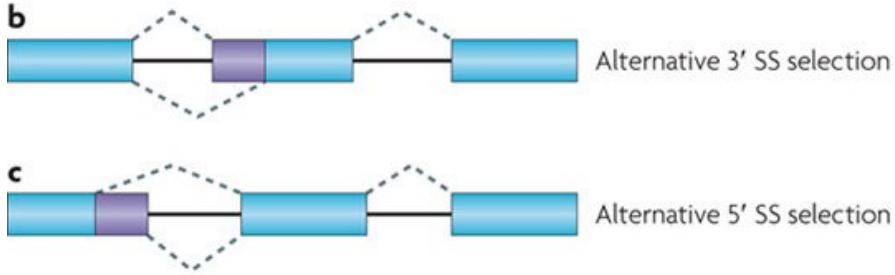

Nature Reviews | Genetics

## Different types of Alternative Splicing




- The first type is exon skipping, in which a type of exon known as cassette exon is spliced out of the transcript together with its flanking introns.

- Exon skipping accounts for nearly 40% of AS events in higher eukaryotes.




- The second and third types are alternative 3’ splice site (3’ SS) and 5’ splice site (5’ SS).

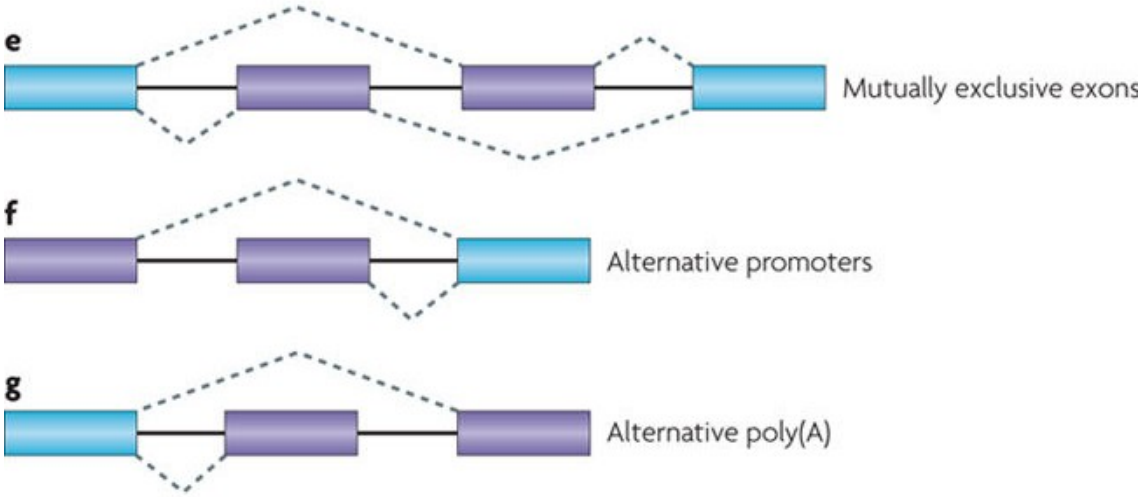
- These types of AS events occur when two or more splice sites are recognized at one end of an exon.

- Alternative 3’ SS and 5’ SS selection account for 18.4% and 7.9% of all AS events in higher eukaryotes, respectively.

## Different types of Alternative Splicing


- The fourth type is intron retention, in which an intron remains in the mature mRNA transcript.

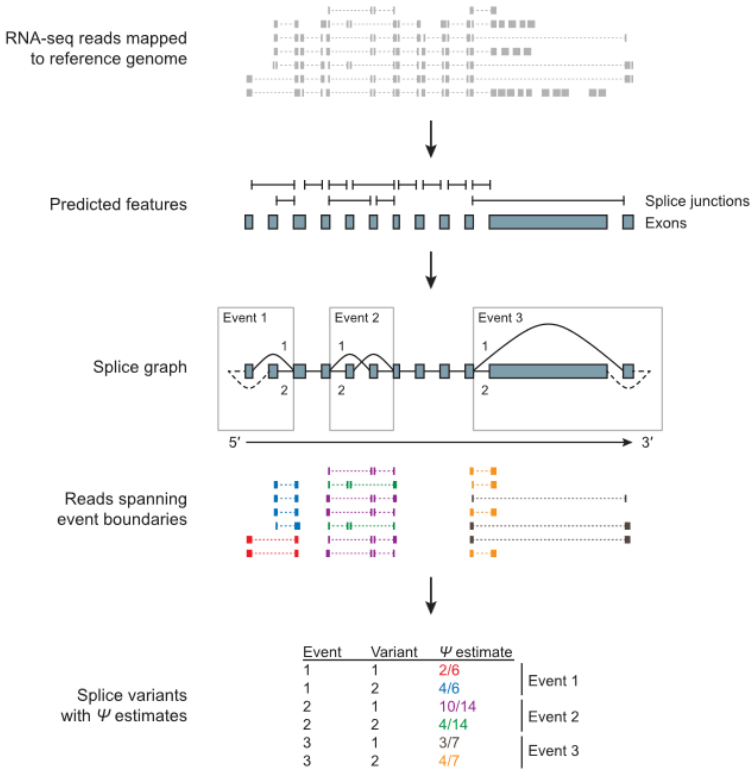
- This is the rarest AS event in vertebrates and invertebrates, accounting for less 5% of known events.




- Less frequent, complex events that give rise to alternative transcripts variant include:
  - ❑ Mutually Exclusive Exons
  - ❑ Alternative Promoter usage
  - ❑ Alternative polyadenylation

# Computational Approaches

- Most available methods for analysis of transcripts variant from RNA-Seq data fall into categories:
  - ❑ Methods for quantification of defined splice events.
  - ❑ Methods for the reconstruction and quantification of full-length transcripts.
  - ❑ The splice events of a gene can be described by a directed acyclic splice graph, where:
    - ✓nodes correspond to transcript starts , end and splice sites
    - ✓edges correspond to exonic regions and splice junctions.




Goldstein et al., PlosOne, 2016

## Computational Approaches

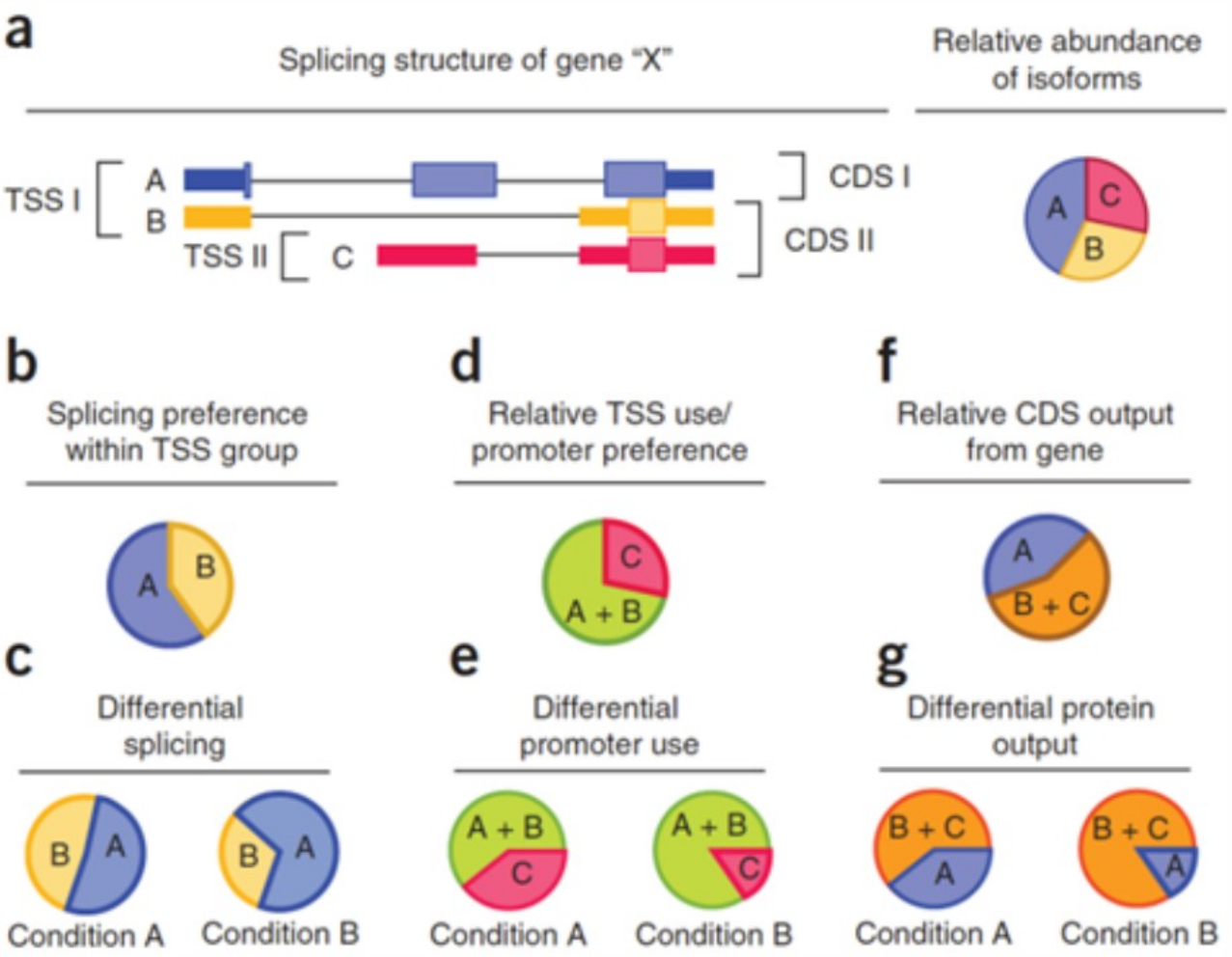
- Recent advances in high-thoroughput technologies have facilitated studies of AS.

- RNA-Seq has become a powerful tool for quantitative profiling of AS.

- Several tools have been implemented:
  - ❑ MISO
  - ❑ SpliceTrap
  - ❑ AlexaSeq
  - ❑ rSeqDiff
- These tools are designed for two samples comparison, but do not handle replicates.

- Other tools are:
  - ❑ **Cufflink**
  - ❑ FDM
  - ❑ DiffSplice
  - ❑ DEXSeq
  - ❑ **rMATS**

**Cufflink - Alternative Splicing through assembly of aligned reads**




Trapnell et al., 2012, Nature Protocol

## MATS (Multivariate Analysis of Transcript Splicing)

- Uses a Bayesian approaches to model between-sample correlation in splicing.
- Uses a simulation-based approach to generate p-values and FDR.




# Gene Fusion

**Dr. Giorgio Giurato, PhD**
ggiurato@unisa.it

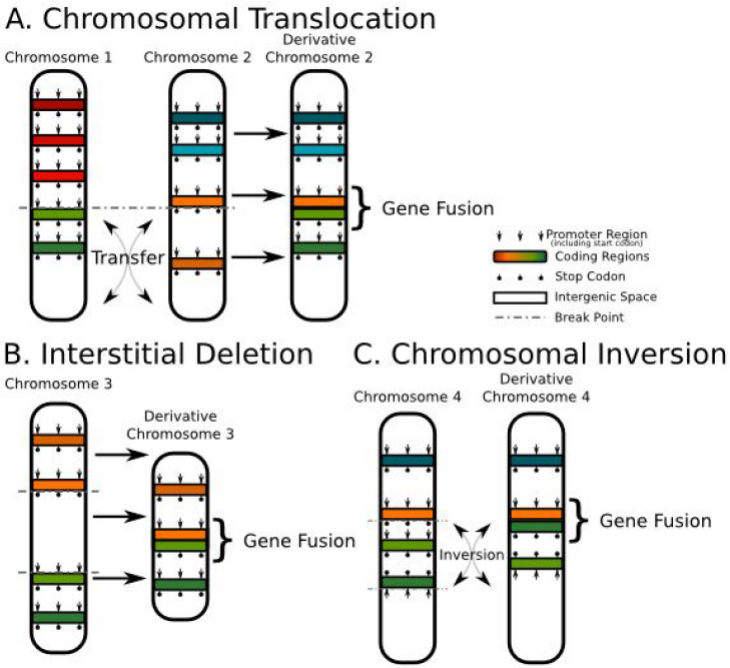
Laboratory of Molecular Medicine and Genomics
Genomix4Life Srl
Department of Medicine, Surgery and Dentistry 'Schola Medica Salernitana'
University of Salerno

## Training Course on Best practise for RNA-Seq data analysis

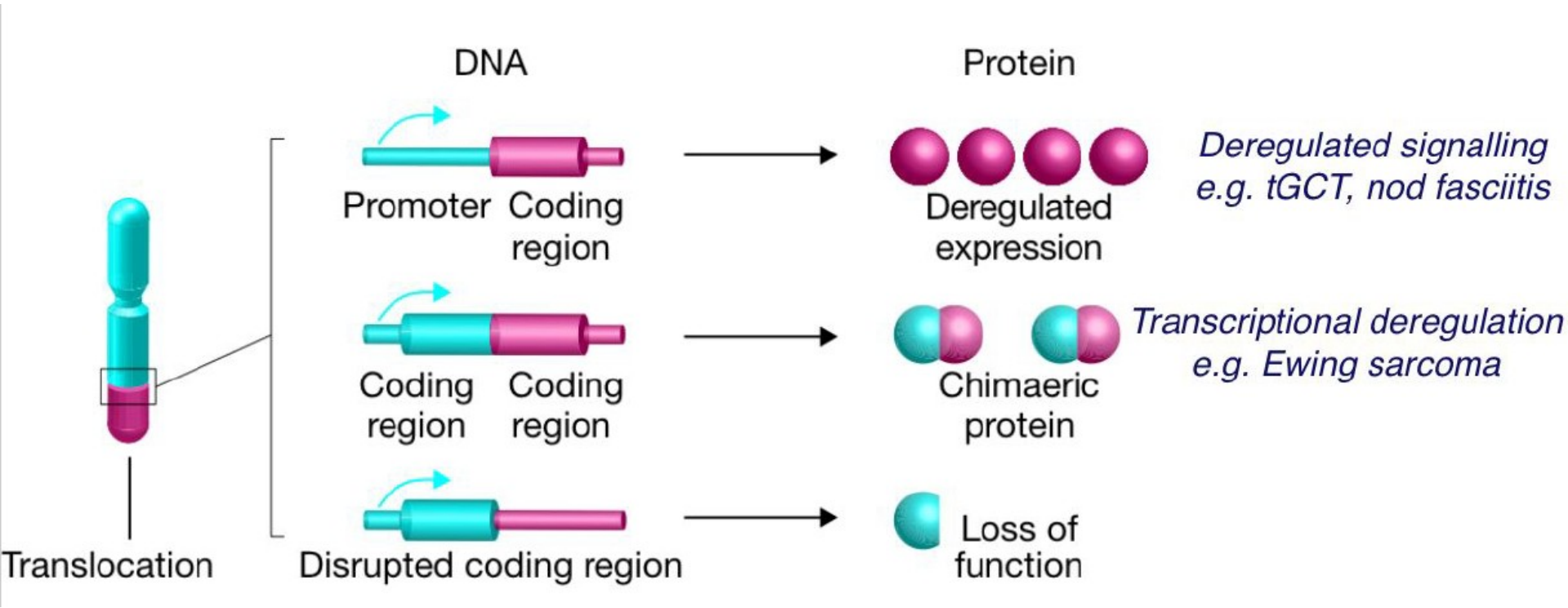
Sept 27-29 2017 – Salerno

# Fusion Genes

- Fusion genes are a prototypical example of pathogenomic mutation.
- A fusion gene is a hybrid gene formed from two previously separate genes.
- It can occurs as a result of traslocation, deletion or chromosomal inversion.
- Detection and characterization of fusion genes has been of great importance for clinical purpose, as well as, for understanding tumorogenesis.



## Effects of fusion proteins




- The abnormal genetic rearrangement contributes to the growth and spread of tumor cells.

- This because the abnormal proteins made by fusion genes appear to be much more active than the normal versions.

## Next-Generation Sequencing

- High-throughput sequencing enables systematic discovery of fusion genes with high sensitivity and precision.

- High-throughput sequencing identifies multiple fusion genes in individual samples, presenting a challenge to distinguish oncogenic “driver” from unimportant “passenger” aberrations.

- The bioinformatic approaches used to identify fusion genes fall into two conceptual classes:

  - ❑ mapping-first approaches: the reads are mapped to genome and gene to identify discordantly mapping reads.

  - ❑ assembly-first approaches: the reads are assembled into longer transcript sequences followed by identifying chimeric transcripts consistent with chromosomal rearrangements.

- Evidence supporting predicted fusions is typically measured by the number of RNA-Seq fragments found as split reads that directly overlap the fusion transcript chimeric junction.

## Bioinformatics tools

- Implementation of several bioinformatic approaches vary in:
  - ❑ Read alignment tools.
  - ❑ Genome database and gene set resources.
  - ❑ Criteria for reporting candidate fusion transcripts and excluding false positive.

- Depending on the fusion prediction tool chosen, a process can taker several days worth of computing and result with a list of hundred of thousands of fusion genes candidates.

- Several tools have been implemented based on different strategies for fusion detections:
  - ❑ Tophat-Fusion
  - ❑ ChimeraScan
  - ❑ FusionCatcher
  - ❑ STAR-Fusion

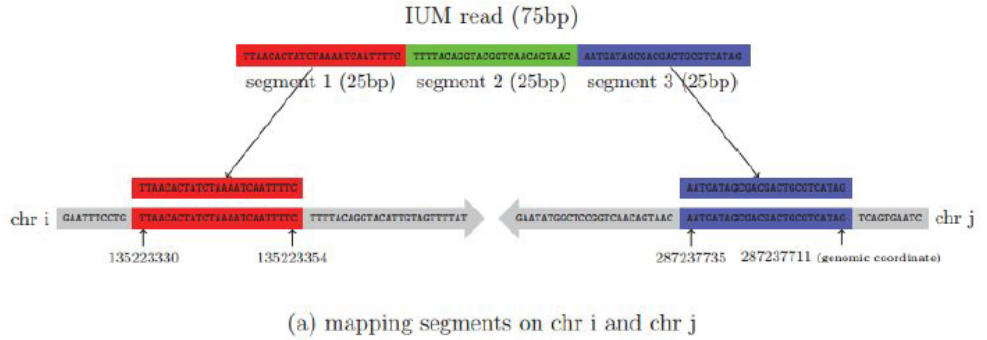
# Tophat-fusion

Bowtie: align reads located entirely in exons

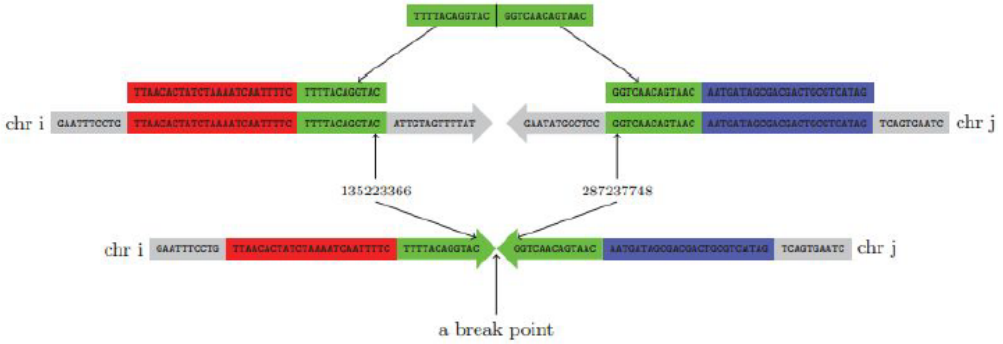
⇩

IUM: Initially unmapped reads split into 25bp

⇒




(a) mapping segments on chr i and chr j




⇦

- Each read must overlap 13bp on both sides of fusion
- Fusion: on different chromosomes or 100,000bp distance
- Penalties for alignments that span:
  - introns (-2)
  - indels (-4)
  - fusions (-4)

\+

- Support of single-end reads
- Many options

\-

- Slow
- Many false positives

# FusionCatcher

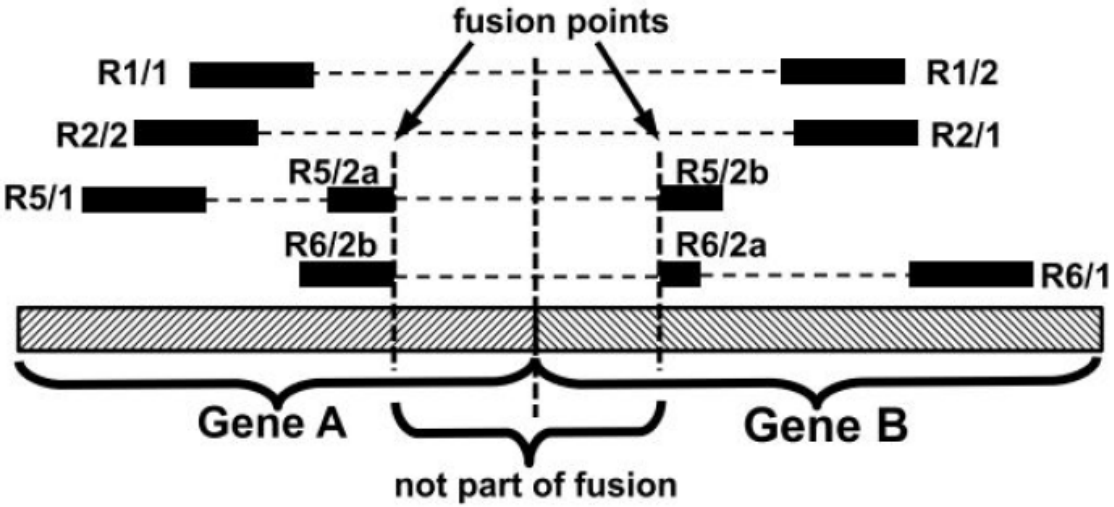
Bowtie maps to genome:
- Unmapped reads are kept
- Candidate fusions: read-pairs mapping to two different genes

⇩

Preliminary fusions filtered for:
- Paralogs
- pseudogenes
- micro/transfer/small-nucleolar RNA
- ConjoinG database
- Previous found in healthy person (Body Map 2.0)

⇩

Unmapped reads aligned with BLAT/STAR/Bowtie2




\+

- Easy to install/run
- High specificity

\-

- Medium sensitivity

- Chimera
- ChimeraScan
- CompleteGenomics
- DeFuse
- EricScript
- FusionCatcher
- FusionMap
- GMAP
- JAFFA
- STAR
- STAR Fusion
- TopHat-Fusion

## FuMa: reporting overlap in RNA-seq detected fusion genes

**Youri Hoogstrate[1,2], René Böttcher[1], Saskia Hiltemann[1,2], Peter J. van der Spek[2], Guido Jenster[1] and Andrew P. Stubbs[2,*]**

[1]Department of Urology and [2]Department of Bioinformatics, Erasmus University Medical Center, Rotterdam, 3000 CA, The Netherlands

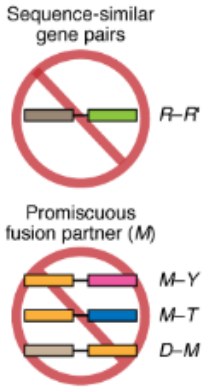
| Left Genes | Right Genes | STAR | TopHat Fusion |
|---|---|---|---|
| FOO1 | BAR1 | UID_A=chr1:12-34 | |
| FOO2 | BAR2 | | TID_A=chr4:66-77 |
| DOX1 | BOX5 | UID_B=chr5:85-95 | TID_B=chr5:88-99 |

## STAR-Fusion

- STAR-Fusion is a largely used tool for fusion genes detection.

- It performs a fast mapping of fusion evidence to reference transcript structure annotation and filters likely artefacts to report accurtate fusion prediction.

- STAR-fusion workflow:
  - ❑ Illumina  RNA-Seq reads are aligned to the genome.
  - ❑ Discordant and split-reads are identified and mapped to reference transcript annotation.
  - ❑ Those reads corresponding to artifacts are filtered.
  - ❑ Fusion candidates containing sequence-similar gene pairs or promiscuous fusion patterns are excluded as likely false positive.

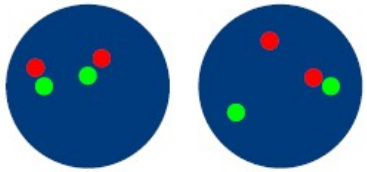


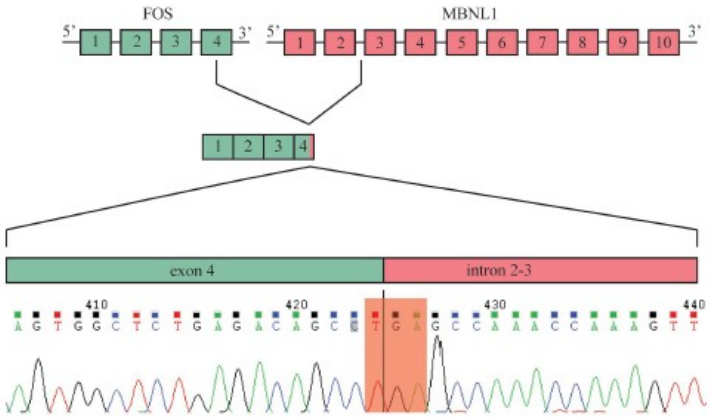
## Cutoff

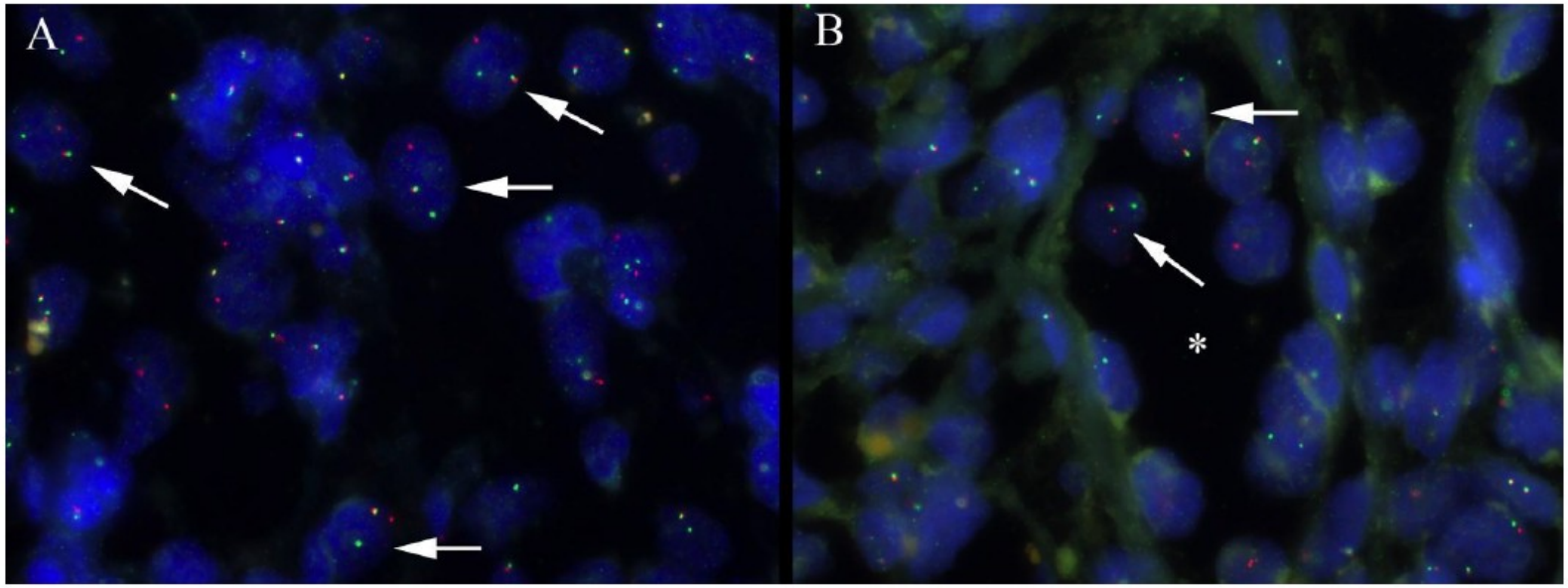
- It is important to exclude false positive fusion genes.
- Gene black list included:
  - ❑ Mitocondrial and ribosomal genes.
  - ❑ Pseudogenes according to the three annotation: Ensembl, ENTREZ Gene Db and HUGO.
- A true fusion junction is likely to present a canonical splice pattern:
  - ❑ GT-AG (~ 98.71%)
  - ❑ GC-AG (~ 0.56%)
  - ❑ AT-AC (~ 0.05%)

# Fusion Validation

- Validation with Sanger Sequencing
- FISH/IHC to identify other cases

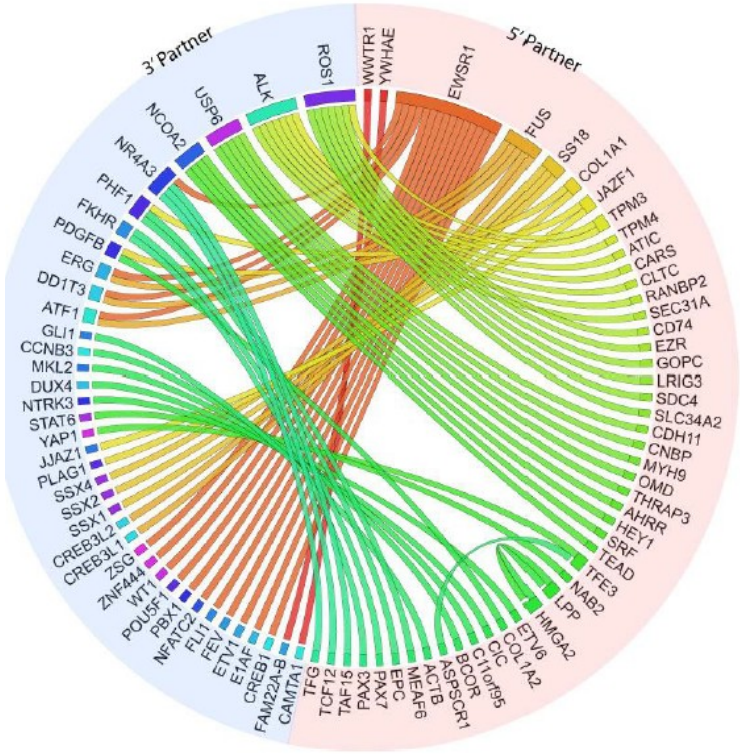





## Conclusion

- Many different tools/strategies for fusion detection:
  - ❑ Discordant reads are analyzed.
  - ❑ Different filter criteria.
  - ❑ Different strategies for fusion boundary detection.
- Combine different tools.



## rMATS command line

- First of all it is necessary to index the genome with STAR:
  - Create on "Scrivania" a folder where save the results.
    - `mkdir rMATS_out`
  - Inside the folder create another sub-folder: STAR_Index
- Move inside `rMATS_out folder (cd rMATS_out)`
  - Here type the command to create the index of the genome. The command to type following:
    - ✓ `STAR --runMode genomeGenerate --genomeDir STAR_Index --genomeFastaFiles /home/studente/Scrivania/Dataset_Corso/Danio_Rerio.cdna.fa --sjdbGTFfile /home/studente/Scrivania/Dataset_Corso/Danio_Rerio.gtf -sjdbOverhang 75`
- Move rMATS folder (Inside Elixir-RNA-Seq-Tools/rMATS.3.2.5 on desktop) and type the following command line:
  - ✓ `python RNASeq-MATS.py -s1 /home/studente/Scrivania/Dataset_Corso/`***`6h1.fastq`***`:/home/studente/Scrivania/Dataset_Corso/`***`6h2.fastq`*** `-s2 /home/studente/Scrivania/Dataset_Corso/`***`2cells_1.fastq`***`: /home/studente/Scrivania/Dataset_Corso/`***`2cells_2.fastq`*** `-gtf /home/studente/Scrivania/Datase_Corso/`***`Danio_rerio.gtf`*** `-bi /home/studente/Scrivania/rMATS_out/`***`STAR_Index`*** `-o /home/studente/Scrivania/rMATS_out/Output_Folder -t paired -len 76`

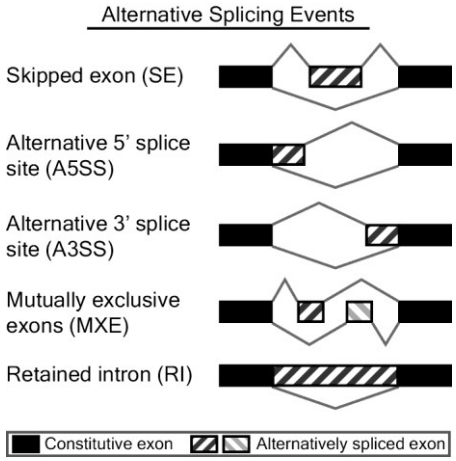
# rMATS output

## Output:

All output files are in outputFolder

- **MATS_output:** A folder that contains rMATS output of AS events. Each output file is sorted by P-values in ascending order.
  - **AS_Event.MATS.JunctionCountOnly.txt** evaluates splicing with only reads that span splicing junctions
    - *IJC_SAMPLE_1:* inclusion junction counts for SAMPLE_1, replicates are separated by comma
    - *SJC_SAMPLE_1:* skipping junction counts for SAMPLE_1, replicates are separated by comma
    - *IJC_SAMPLE_2:* inclusion junction counts for SAMPLE_2, replicates are separated by comma
    - *SJC_SAMPLE_2:* skipping junction counts for SAMPLE_2, replicates are separated by comma

  - **AS_Event.MATS.ReadsOnTargetAndJunctionCounts.txt** evaluates splicing with reads that span splicing junctions and reads on target (
    - *IC_SAMPLE_1:* inclusion counts for SAMPLE_1, replicates are separated by comma
    - *SC_SAMPLE_1:* skipping counts for SAMPLE_1, replicates are separated by comma
    - *IC_SAMPLE_2:* inclusion counts for SAMPLE_2, replicates are separated by comma
    - *SC_SAMPLE_2:* skipping counts for SAMPLE_2, replicates are separated by comma



# rMATS output

- **summary.txt:** A file that contains summary of statistically significant AS events and the identity of each replicate
- **ASEvents:** A folder that contains all possible alternative splicing (AS) events derived from GTF and RNA
- **SAMPLE_1/REP_N:** A folder that contains mapping results of sample_1, replicate N
  - *accepted_hits.bam* is the original tophat output containing both multi-mapped and uniquely mappable reads.
  - *unique.S1.sam* contains uniquely mappable reads only. rMATS uses uniquely mappable reads.
- **SAMPLE_2/REP_N:** A folder that contains mapping results of sample_2, replicate N
  - *accepted_hits.bam* is the original tophat output containing both multi-mapped and uniquely mappable reads.
  - *unique.S2.sam* contains uniquely mappable reads only. rMATS uses uniquely mappable reads.
- **commands.txt:** A list of key commands executed
- **log.RNASeq-MATS:** Log file for running rMATS pipeline

| C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| geneSymb | chr | strand | exonStart | exonEnd | upstream | upstream | downstre | downstre | ID | IJC_SAMPLE_1 | SJC_SAMF | IJC_SAMP | SJC_SAMF | IncFormLe | SkipForml | PValue | FDR | IncLevel1 | IncLevel2 | IncLevelDifference |
| ARHGAP4 | 17 | + | 12877405 | 12877627 | 12862033 | 12862214 | 12883374 | 12883550 | 62650 | 0,4,1 | 0,1,2 | 2,2,1 | 2,6,0 | 172 | 86 | 9.06E-06 | 0.0001 | NA,0.667, | 0.333,0.143,1.0 | -0.059 |
| BRD8 | 5 | - | 1.38E+08 | 1.38E+08 | 1.38E+08 | 1.38E+08 | 1.38E+08 | 1.38E+08 | 17025 | 89,74,86 | 0,0,0 | 68,66,86 | 1,0,1 | 172 | 86 | 9.08E-06 | 0.000101 | 1.0,1.0,1.0 | 0.971,1.0,0.977 | 0.017 |
| WWOX | 16 | + | 78142319 | 78142384 | 78133630 | 78133782 | 78143674 | 78143732 | 31848 | 184,175,172 | 4,3,1 | ######## | 4,0,0 | 81 | 51 | 9.08E-06 | 0.000101 | 0.967,0.97 | 0.983,1.0,1.0 | -0.017 |
| SOD2 | 6 | - | 1.6E+08 | 1.6E+08 | 1.6E+08 | 1.6E+08 | 1.6E+08 | 1.6E+08 | 36120 | 254,260,325 | 7,8,21 | ######## | 4,3,8 | 172 | 86 | 9.12E-06 | 0.000101 | 0.948,0.94 | 0.956,0.963,0.915 | -0.019 |

## STAR-Fusion

- The data for our Practical course are inside the folder Dataset_Corse/Gene_Fusion on "Scrivania".
- Here we found twol folders:
  - FASTQ
  - Genome

- Move inside Gene_Fusion folder. Here we will index the genome typing the following command line:
  - perl /usr/bin/FusionFilter/prep_genome_lib.pl -genome_fa /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Genome/Dati/Chr17.fa -gtf /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Genome/Dati/ref_anno_chr17.gtf --blast_pairs /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Genome/Dati/blast_pairs.gene_syms.outfmt6.gz -fusion_annot_lib /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Genome/Dati/fusion_lib.dat.gz
  - perl /usr/bin/FusionFilter/util/index_pfam_domain_info.pl -pfam_domains /home/studenti/Scrivania/Dataset_Corso/Gene_Fusion/Genome/Dati/PFAM.domtblout.dat.gz -genome_lib_dir /home/studente/Scrivania/GeneFusion/ctat_genome_lib_build_dir

## STAR-Fusion

- Now we can run STAR-Fusion with standard parameter:
- STAR-Fusion **--genome_lib_dir** ctat_genome_lib_build_dir **--left_fq** /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Fastq/BT474-Fisubset.Left.fq **--right_fq** /home/studente/Scrivania/Dataset_Corso/Gene_Fusion/Fastq/BT474-Fisubset.Rigth.fq **--output_dir** Out

## Output_file

| A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| #FusionName | JunctionReadCount | SpanningFragCount | SpliceType | LeftGene | LeftBreakpoint | RightGene | htBreakpo | LargeAnchorSupport | LeftBreakDinuc | BreakEntr | htBreakD | itBreakEnt | FFPM |
| THRA--THRA1/BTR | 27 | 93 | ONLY_REF_SPLICE | THRA^ENSG00000126351.12 | chr17:40086853:+ | THRA1/BTR^ENSG00000235300.4 | 17:482943 | YES_LDAS | GT | 1.8892 | AG | 1.9656 | 23875.85 |
| THRA--THRA1/BTR | 5 | 93 | ONLY_REF_SPLICE | THRA^ENSG00000126351.12 | chr17:40086853:+ | THRA1/BTR^ENSG00000235300.4 | 17:483073 | YES_LDAS | GT | 1.8892 | AG | 1.4295 | 19498.61 |
| ACACA--STAC2 | 12 | 52 | ONLY_REF_SPLICE | ACACA^ENSG00000278540.4 | chr17:37122531:- | STAC2^ENSG00000141750.6 | 17:392181 | YES_LDAS | GT | 1.9656 | AG | 1.9656 | 12733.78 |
| RPS6KB1--SNF8 | 10 | 43 | ONLY_REF_SPLICE | RPS6KB1^ENSG00000108443.13 | chr17:59893325:+ | SNF8^ENSG00000159210.9 | 17:489439 | YES_LDAS | GT | 1.3753 | AG | 1.8323 | 10545.17 |
| TOB1--SYNRG | 8 | 30 | ONLY_REF_SPLICE | TOB1^ENSG00000141232.4 | chr17:50866058:- | SYNRG^ENSG00000275066.4 | 17:375206 | YES_LDAS | GT | 1.4566 | AG | 1.8892 | 7560.684 |
| VAPB--IKZF3 | 4 | 46 | ONLY_REF_SPLICE | VAPB^ENSG00000124164.15 | chr20:58389517:+ | IKZF3^ENSG00000161405.16 | 17:397777 | YES_LDAS | GT | 1.9656 | AG | 1.7819 | 9948.269 |
| ZMYND8--CEP250 | 2 | 44 | ONLY_REF_SPLICE | ZMYND8^ENSG00000101040.19 | chr20:47224317:- | CEP250^ENSG00000126001.15 | 20:354906 | NO_LDAS | GT | 1.8295 | AG | 1.8062 | 9152.408 |

- **LargeAnchorSupport**: indicates whether there are split reads that provide 'long' alignments on both sides of the putative breakpoint.
- **FFPM**: Normalized measures of the split reads and spanning fragments