

# Bioinformatics

## SplitThreader

MSc. Vicente Machaca Arceda

Universidad Nacional de San Agustín

2021

# Content

## 1 Introduction

- Concepts
- Problem
- Research question

## 2 SplitThreader

- Proposal
- Pipeline
- Results
- Conclusions

# Overview

## 1 Introduction

- **Concepts**

- Problem

- Research question

## 2 SplitThreader

- Proposal

- Pipeline

- Results

- Conclusions

# DNA

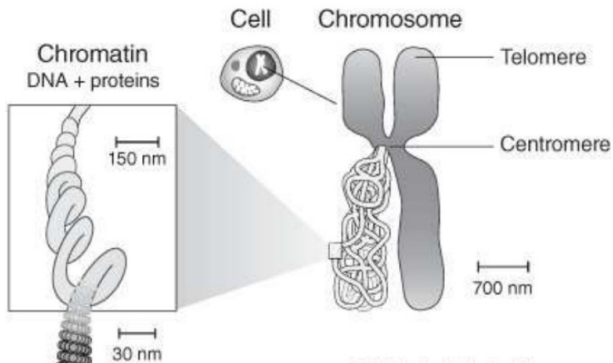


Figure: Location of DNA. Source: [1]

# Genes

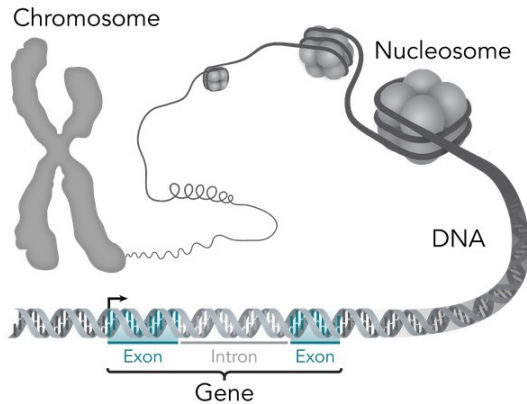


Figure: DNA and genes. Source: [1]

# Transcription and translation

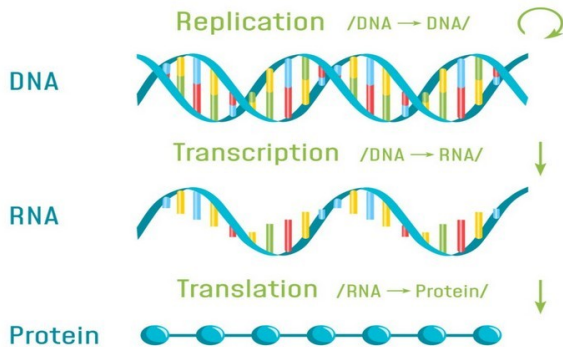
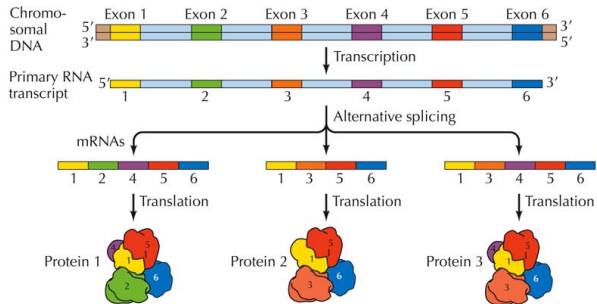


Figure: DNA and genes. Source: [2]

# Alternative splicing



THE CELL, Fourth Edition, Figure 5.5 © 2006 ASM Press and Sinauer Associates, Inc.

Figure: Alternative splicing. Source: [3].

# DNA example

```
>gb:MN988668|Organism:Wuhan seafood market pneumonia virus|Strain  
Name:2019-nCoV_WHU01|Segment:null|Host:Human  
TTAAAGGTTTATACCTTCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAAC  
GAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACCTCACGCAGTATAATTAATAACT  
AATTACTGTCGTTGACAGGACACGAGTAACCTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGT  
TGCAGCCGATCATCAGCACATCTAGGTTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGCC  
CTGGTTTTCAACGAGAAAAACACACGTCCAACCTCAGTTTTGCCTGTTTTACAGGTTTCGCGACGTGCTCGTACG  
TGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGGC  
TTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTTCATCAAACGTTCCGATG  
CTCGAACTGCACCTCATGGTCATGTTATGTTGAGCTGGTAGCAGAACTCGAAGGCATTAGTACGGTCG  
TAGTGGTGAGACACTTGGTGTCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCAGCAAGGTTCTT  
CTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAG  
GCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGT  
TACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGC  
CCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTCATGCACTTTGT  
CCGAACAACCTGGACTTTATTGACACTAAGAGGGGTGTACTGCTGCCGTGAACATGAGCATGAAATTGC  
TTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCAGACACCTTTGAAATTAATTTGGCAAAGAAA  
TTTGACACCTTCAATGGGGAATGTCCAAATTTTGTATTTCCCTTAAATTCATAATCAAGACTATTCAAC  
CAAGGGTTGAAAAGAAAAAGCTTGATGGCTTTATGGGTAGAATTGATCTGTCTATCCAGTTGCGTCACC
```

Figure: A piece of COVID-19 DNA.



# Genomics and Big Data

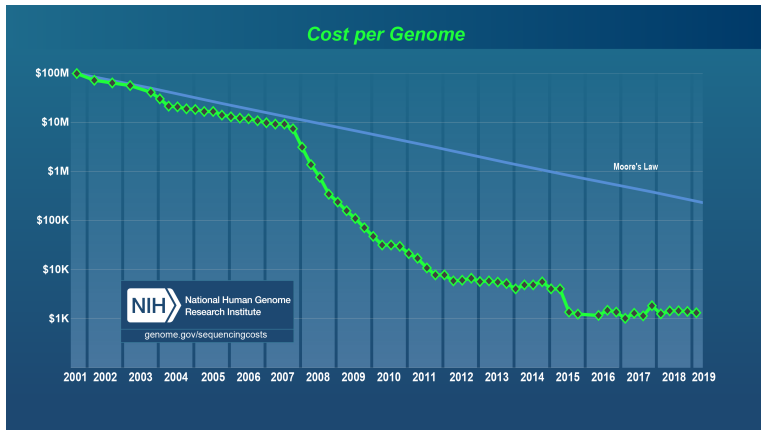


Figure: Cost per genome sequencing over the years.

# Genomics and Big Data

In 2009, genomics data reached about 0.8 ZB. Moreover in 2020 they reached about 40 ZB [4]

# Genomics and Big Data

- 6.4 billions of bases.
- 20k genes approximately.
- No technology exist that can read an entire chromosome from end to end.
- Some changes in the genome encode normal variation like hair color, other can cause diseases.

# Structural variants

Single Nucleotide Variant



Deletion



Insertion



Tandem Duplication



Interspersed Duplication



Inversion



Translocation



Copy Number Variant



## Types of Variants

Figure: Example of structural variants. Source: [5]

# Copy number variation

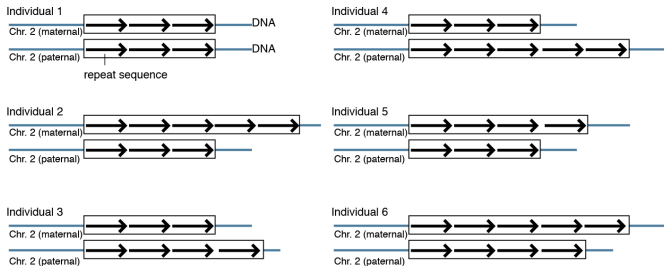


Figure: Example of copy number variation. Source: [6]

# Gene fusion

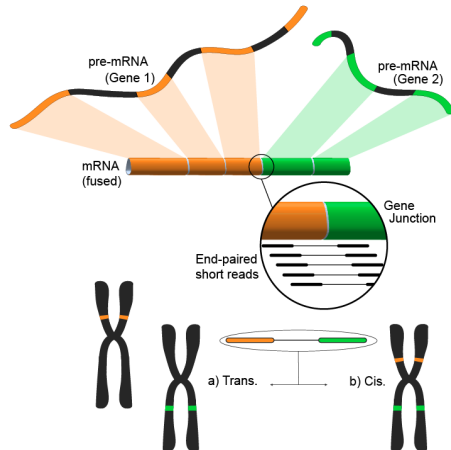


Figure: Gene fusion example.

# Gene fusion

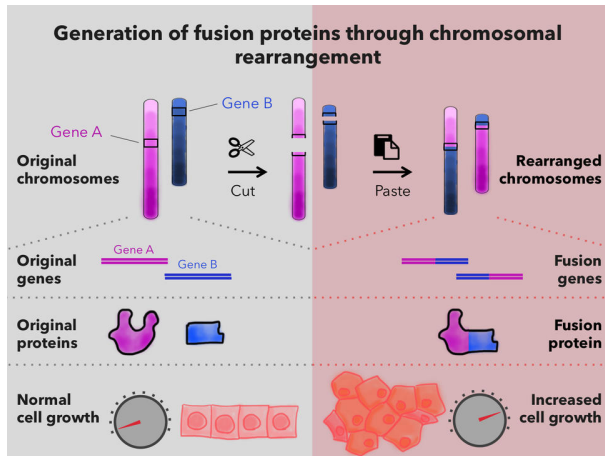


Figure: Gene fusion example.

# Chromosome-scale rearrangements



Figure: 46 Chromosomes presented in cells.



# Chromosome-scale rearrangements

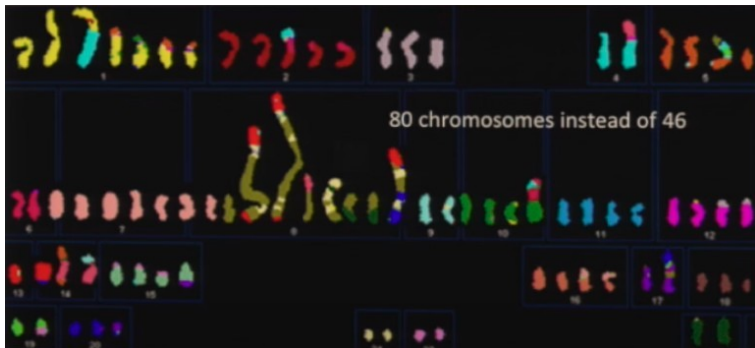


Figure: Cell line from a woman with metastatic breast cancer in 1971. Tumor cells have been grown and studied in the lab ever since.

# Overview

- 1 Introduction
  - Concepts
  - **Problem**
  - Research question

- 2 SplitThreader
  - Proposal
  - Pipeline
  - Results
  - Conclusions

# Problem

Genomic instability is one of the **hallmarks of cancer** [7, 8], resulting in:

- Widespread copy number changes.
- Structural variants.
- Chromosome-scale rearrangements.

**Copy number variants** and **gene fusions** are common drivers in cancer [9, 10].

# Problem

The available algorithms for identifying gene fusions **do not have perfect specificity** (false positive rate). Require a joint analysis of genomic and transcriptomic data to correctly analyze.

**Rearrangements variants are difficult to study**, because of the sheer complexity of rearrangements, which often include adjacencies between distant regions of a chromosome or even between unrelated chromosomes

# Overview

- 1 Introduction
  - Concepts
  - Problem
  - **Research question**
- 2 SplitThreader
  - Proposal
  - Pipeline
  - Results
  - Conclusions

# Research question

Exploration and analysis of rearrangements in cancer genomes could be performed with a web platform?

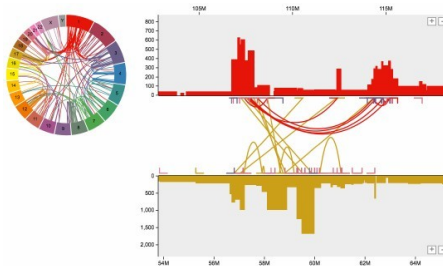
# Overview

- 1 Introduction
  - Concepts
  - Problem
  - Research question
- 2 SplitThreader
  - **Proposal**
  - Pipeline
  - Results
  - Conclusions

# SplitThreader

## Proposal

SplitThreader, an open source interactive **web application** for analysis and visualization of genomic rearrangements and copy number variation in cancer genomes [11].





# Overview

## 1 Introduction

- Concepts
- Problem
- Research question

## 2 SplitThreader

- Proposal
- **Pipeline**
- Results
- Conclusions

# SplitThreader pipeline

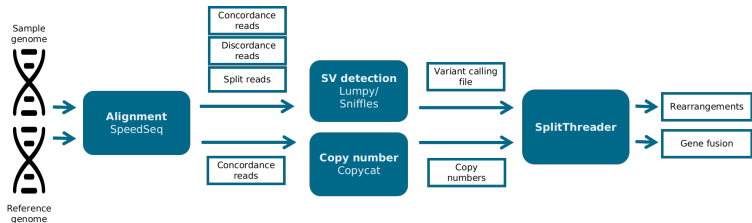


Figure: SplitThreader pipeline.

# Variant Calling File

chrom1	start1	stop1	chrom2	start2	stop2	name	type	split
1	17051740	17051740	1	234912188	234912188	35665	BND	71
1	47659735	47659735	8	105739138	105739138	573599	BND	6
1	87069066	87069066	2	82854932	82854932	571553	BND	6
1	109650635	109650635	22	30163373	30163373	575755	BND	36
1	150593722	150593722	5	55447995	55447995	572639	BND	19
1	153306043	153306043	16	76228788	76228788	575219	BND	6
1	168186186	168186186	1	182274316	182274316	20968	BND	11
1	201288206	201288206	10	52642286	52642286	574013	BND	9
1	208992122	208992122	3	87327147	87327147	572038	BND	
...								

Figure: Example of a Variant Calling File.

# Copy Number Profile

## Method

chromosome	start	end	coverage
1	0	10000	0
1	10000	20000	0.9605
1	20000	30000	0
1	30000	40000	0
1	40000	50000	0.0059
1	50000	60000	0.775
1	60000	70000	0.6154
1	100000	110000	0.3666
...			

Figure: Example of a Copy Number Profile.

# Overview

- 1 Introduction
  - Concepts
  - Problem
  - Research question
- 2 SplitThreader
  - Proposal
  - Pipeline
  - **Results**
  - Conclusions

# SplitThreader

## Results

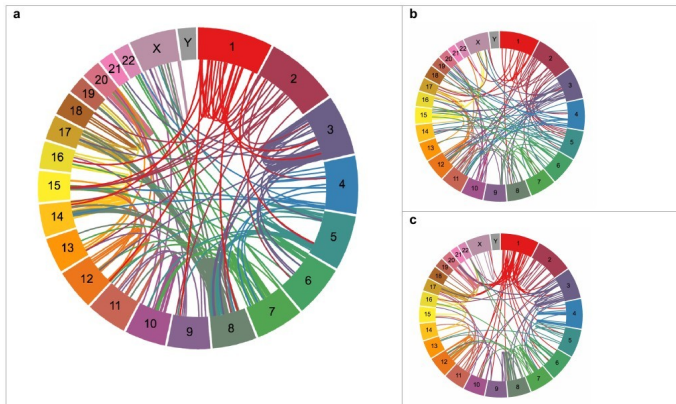


Figure: Circos plots showing genomic rearrangements in the cell lines SK-BR-3 (a), A549 (b), and MCF-7 (c).

# SplitThreader

## Results

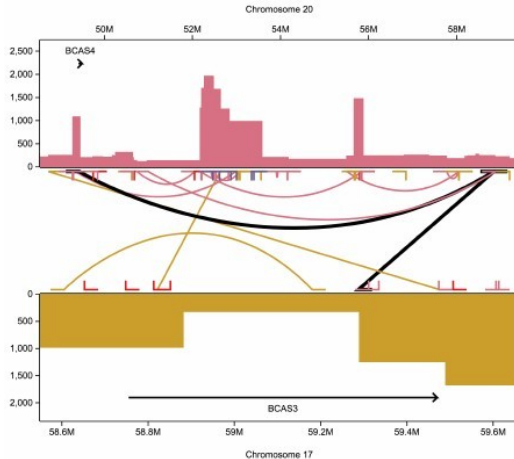


Figure: BCAS4-BCAS3 two-hop gene fusion gene fusion in MCF-7.

# SplitThreader

## Results

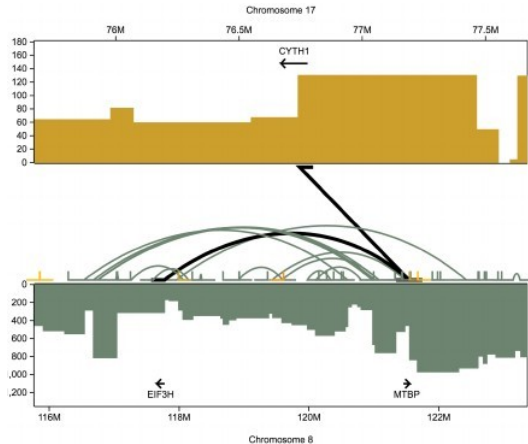


Figure: CPNE1-PHF20-PREX1 two-hop gene fusion in SK-BR-3.



# Overview

- 1 Introduction
  - Concepts
  - Problem
  - Research question
- 2 SplitThreader
  - Proposal
  - Pipeline
  - Results
  - **Conclusions**

# Conclusions

Structural variant are key markers in cancer genomics. Most of them are related to high copy number variants and gene fusions. Nevertheless, it is difficult to detect this variants.

# Conclusions

Structural variant are key markers in cancer genomics. Most of them are related to high copy number variants and gene fusions. Nevertheless, it is difficult to detect this variants.

Visualization is a emerging area applied in Bioinformatics, it is used in Proteomics, Genomics, Metagenomics and Cancer genomics. SplitThreader and MoMI-G help the analysis and visualization of structural variants.


# Conclusions

Structural variant are key markers in cancer genomics. Most of them are related to high copy number variants and gene fusions. Nevertheless, it is difficult to detect this variants.

Visualization is a emerging area applied in Bioinformatics, it is used in Proteomics, Genomics, Metagenomics and Cancer genomics. SplitThreader and MoMI-G help the analysis and visualization of structural variants.

SplitThreader uses a breath-first search algorithm in order to detect gene fusions, then it plots copy number variants and gene fusions. Nevertheless, the results of SplitThreader depends on Lumpy and Sniffles.

# References I

-  J. M. Archibald, *Genomics: A Very Short Introduction*. Oxford University Press, 2018, vol. 559.
-  J. Xiong, *Essential bioinformatics*. Cambridge University Press, 2006.
-  G. BIO, “Gen. bio,”  
<https://sites.google.com/site/bio1040genbio2/home>, 2020,  
accessed: 2020-03-20.
-  A. Prabahar and S. Swaminathan, “Perspectives of machine learning techniques in big data mining of cancer,” in *Big Data Analytics in Genomics*. Springer, 2016, pp. 317–336.

# References II



PacBio, “Two review articles assess structural variation in human genomes,” <https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes> 2021, accessed: 2021-05-07. [Online]. Available: <https://www.pacb.com/blog/two-review-articles-assess-structural-variation-in-human-genomes>







N. H. Genome, “Copy number variation (cnv),” <https://www.genome.gov/genetics-glossary/Copy-Number-Variation>, 2021, accessed: 2021-05-07. [Online]. Available: <https://www.genome.gov/genetics-glossary/Copy-Number-Variation>



D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *cell*, vol. 144, no. 5, pp. 646–674, 2011.

# References III

-  P. J. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, “Mechanisms of change in gene copy number,” *Nature Reviews Genetics*, vol. 10, no. 8, pp. 551–564, 2009.
-  A. Shlien and D. Malkin, “Copy number variations and cancer,” *Genome medicine*, vol. 1, no. 6, pp. 1–9, 2009.
-  F. Mitelman, B. Johansson, and F. Mertens, “The impact of translocations and gene fusions on cancer causation,” *Nature Reviews Cancer*, vol. 7, no. 4, pp. 233–245, 2007.
-  M. Nattestad, M. C. Alford, F. J. Sedlazeck, and M. C. Schatz, “Splitthreder: Exploration and analysis of rearrangements in cancer genomes,” *bioRxiv*, p. 087981, 2016.

# Questions?

