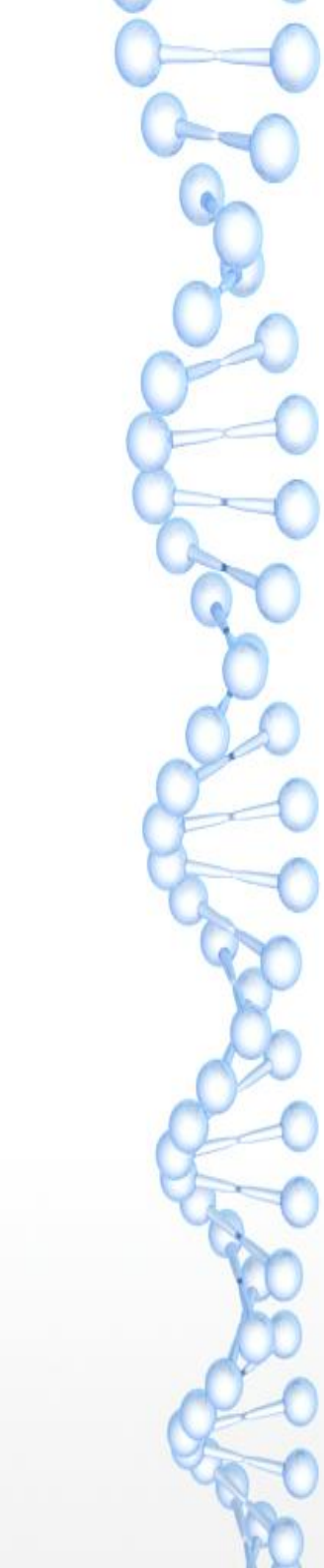


# Light-weight Visualization System for Large Genomics Data

# DNA Sequencing

**Human Genome Project (HGP) was an international research effort to determine the sequence of the human genome and identify the genes that it contains.**

**The HGP was a 13 years (1990 – 2003) effort that cost billions of effort coordinated by:**

**1- National Institutes of Health (NIH)**

**2- Department of Energy (DOE)**

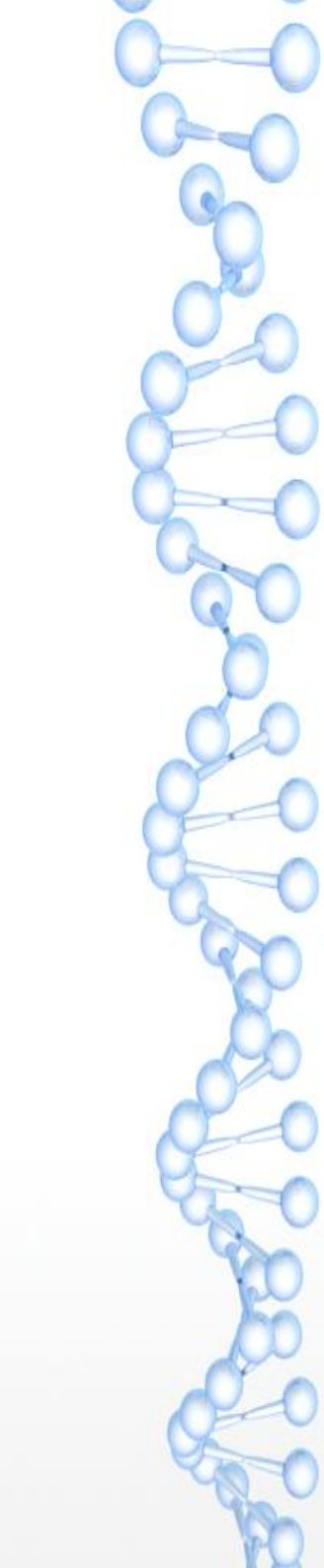
# Applications of HGP

- **Identification of human genes and their functions**
- **Understanding of polygenetic disorders such as cancer**
- **Improvements in gene therapy**
- **Improved diagnosis of diseases**
- **Development of pharmacogenesis**

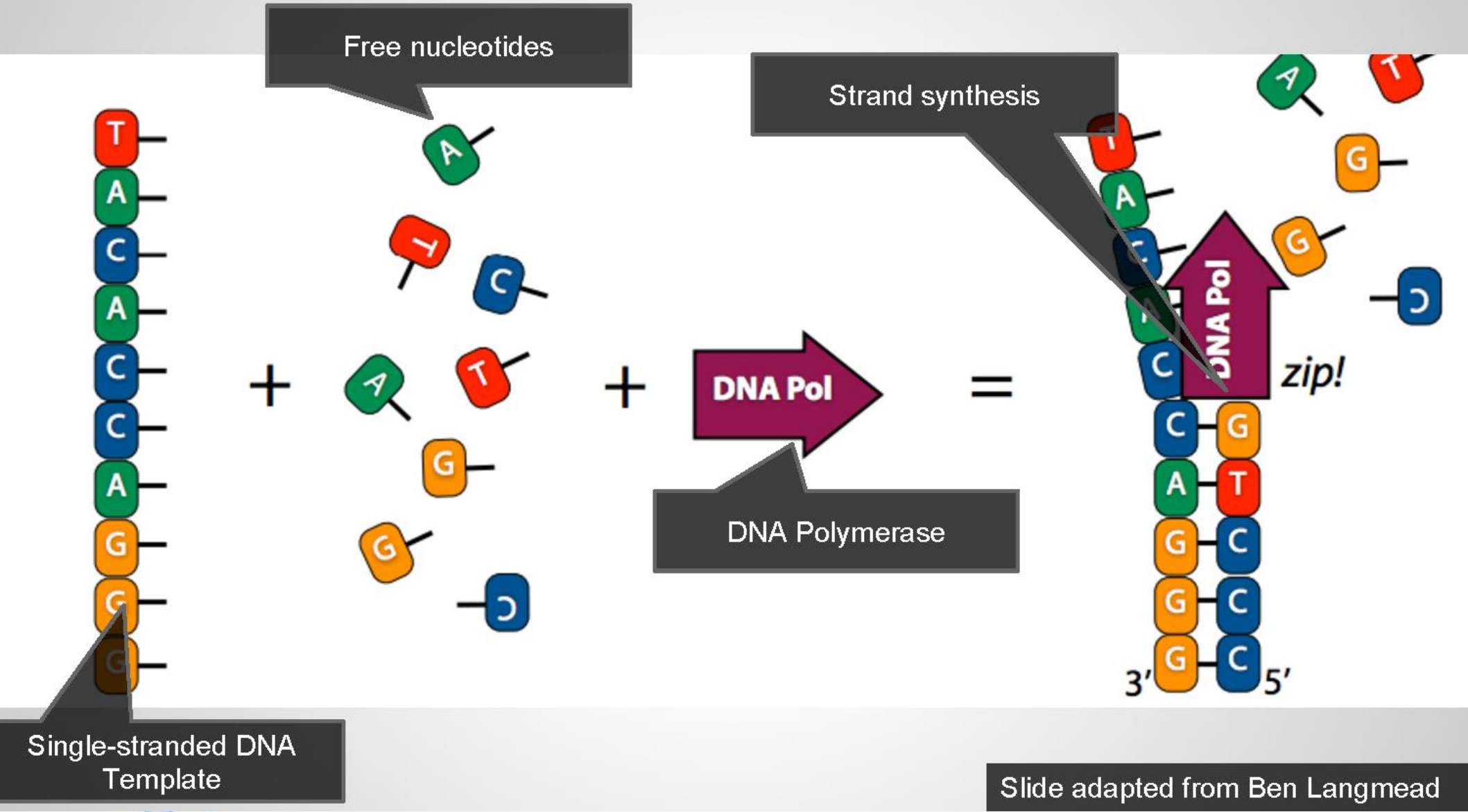
# First-generation sequencing versus Next-Generation Sequencing

Next generation sequencing (NGS), massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionized genomic research. Using NGS an entire human genome can be sequenced within a single day at a cost of 1000 USD.

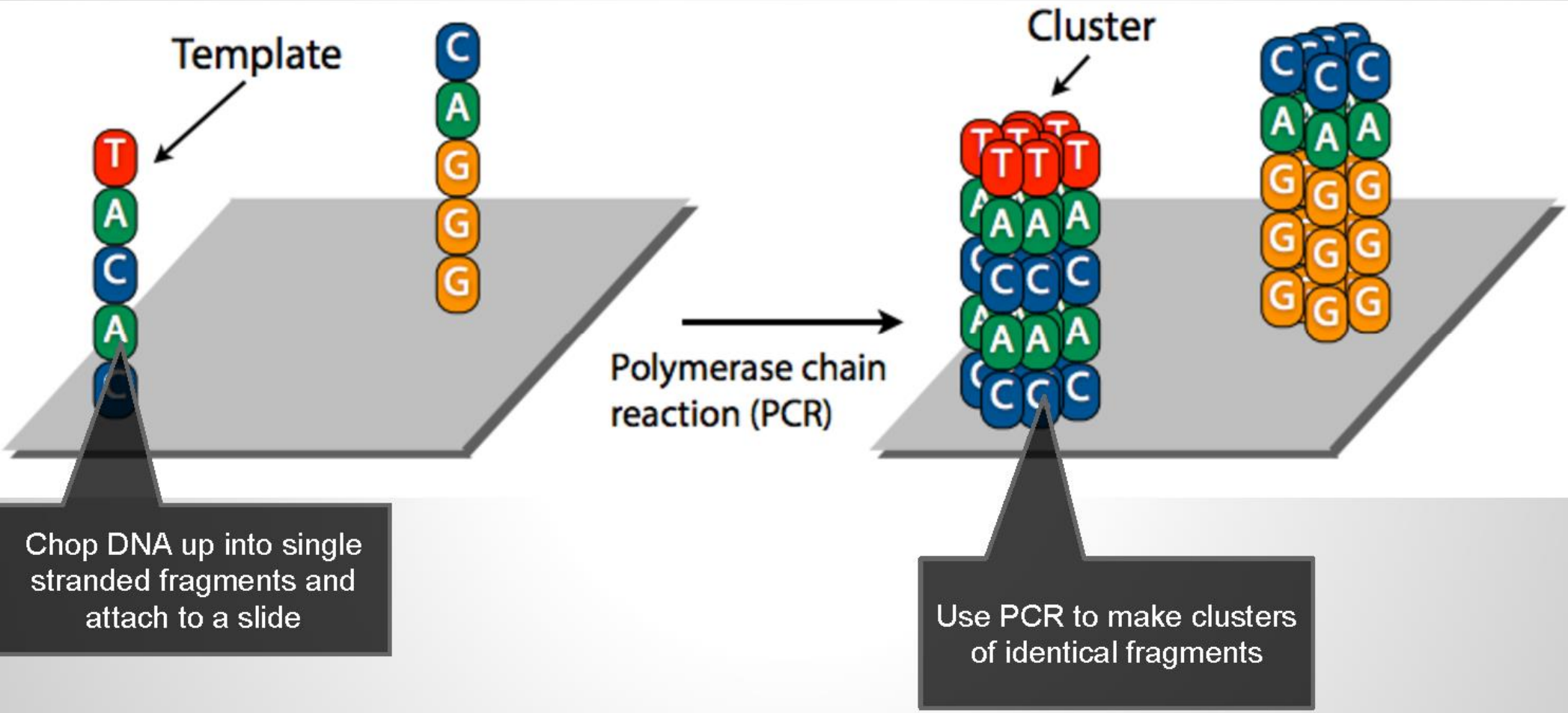
In contrast, the previous Sanger sequencing technology, used to decipher the human genome, required over a decade to deliver the final draft.



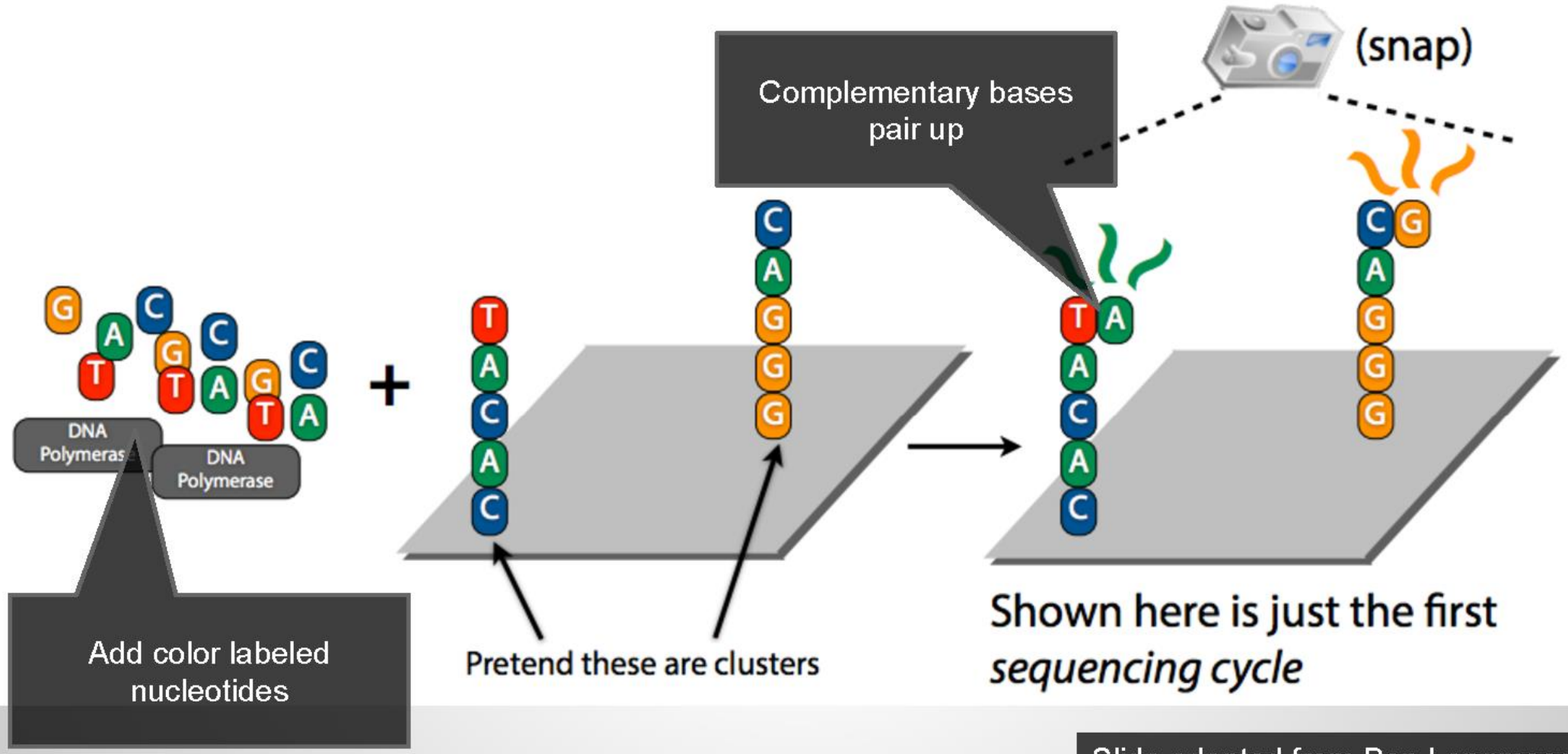
# How does Next-Generation Sequencing take place?



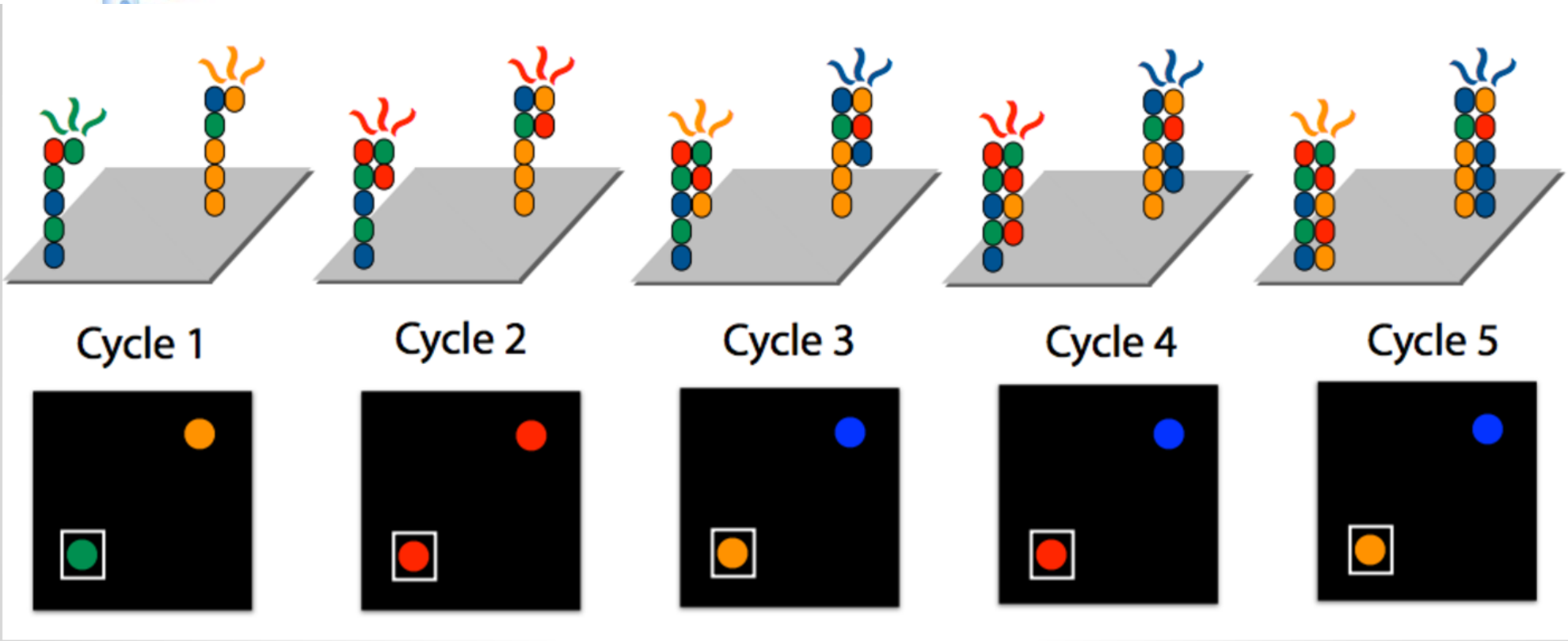

Slide adapted from Ben Langmead

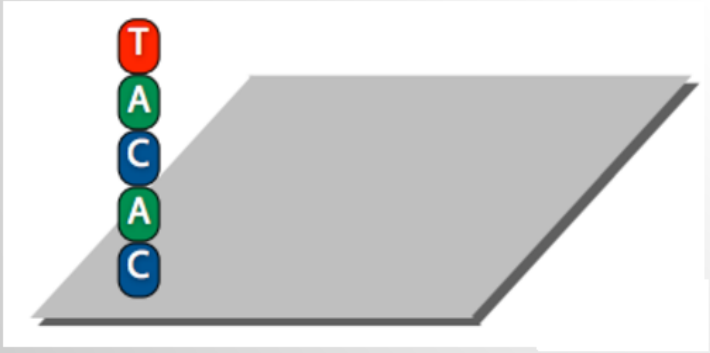



Slide adapted from Ben Langmead




Slide adapted from Ben Langmead




Slide adapted from Ben Langmead

# Definition Problem

**Functional fast and low cost data visualization systems of genomic data has been controversial because of their expense in terms of money and computational cost.**

**By rethinking our approach to management of NGS data by visualizing them through various aspects in comparison to a reference genome to spot changes, we can offer best practices for best adoption of these new technologies for genomic data.**

# Motivation

- Batch Mode Generation of quick images.
- Plots associated to databases with annotations.

# Why do we need visualization?

- Detect Mutations
- Assure sequencing covers all exons
- Assure sequencing generates reads covering whole reference genome

# Related Works

**IGV**

- Java based
- loads BAM files, annotations, and other tracks
- Not the fastest in performance

**Tablet**

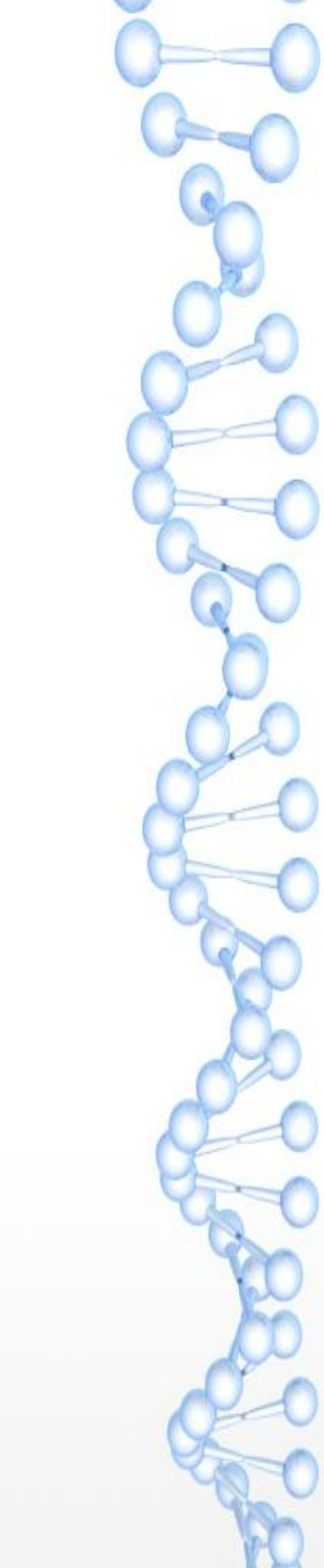
- Delicate balance between performance, features, and aesthetics
- Good interface
- Does not display read insertions correctly

**BamView**

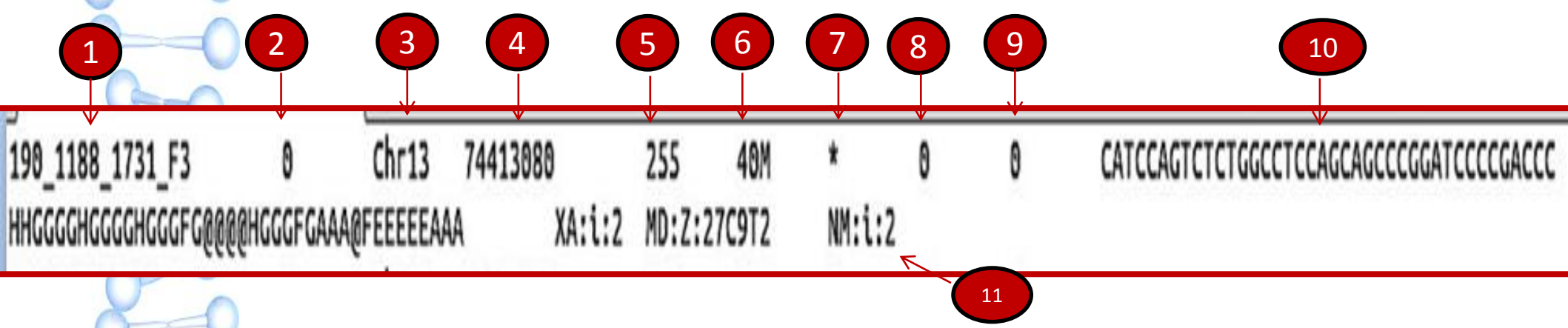
- Very fast simple BAM file viewer
- No much features, just display and drag

# Database Generation

| | |
|---|---|
| 1 | • DNA extraction from a sample |
| 2 | • DNA sequencing |
| 3 | • Raw sequencing reads are aligned to a reference genome |
| 4 | • Convert aligned reads from Sam file format to Bam |
| 5 | • Sort and Index BAM file |



# SAM FILE




**(1) Name Read**

**(2) Flags: Each number in the flags indicate certain properties for Reads like : read paired, read mapped in proper pair, mate reverse strand, first in pair.**

**(3)Chromosome Name: Reference sequence NAME**

**(4)Position : position or start point of the read in alignment to the reference genome**

**(5)MAPQ(Mapping Quality)**

**(6)CIGAR String**

**(7) R NEXT**

**(8) P NEXT**

**(9) Template Length**

**(10) SEQUENCE**

**(11) Tags**

```
190_1188_1731_F3	0	Chr13	74413080	255	40M	*	0	0	CATCCAGTCTCTGGCCTCCAGCAGCCCGGATCCCCGACCC
HHGGGGHGGGGHGGGFG@@@@HGGGFGAAA@FEEEEEAAA	XA:i:2	MD:Z:27C9T2	NM:i:2
689_1478_1241_F3	0	Chr13	74413128	250	75M	*	0	0
GCAGGTGATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACT	*	Z0:i:25149	Z1:i:25149	CM:i:0	NM:i:3
AS:i:641	CS:Z:T131201123203222031321113111001001300033321010002001212203001332210302020212
XX:Z:GCAGGTGATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACT
322_1007_1216_F3	0	Chr13	74413129	250	75M	*	0	0
CAGGTGATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTT	*	Z0:i:25149	Z1:i:25149	CM:i:0	NM:i:3
AS:i:641	CS:Z:T212011232032220313211131110010013000333210100020012122030013322103020202120
XX:Z:CAGGTGATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTT
74_444_1049_F3	16	Chr13	74413134	250	75M	*	0	0
GCCCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCA	*	Z0:i:25141	Z1:i:25141	CM:i:0	NM:i:3
AS:i:641	CS:Z:T0100310212020203012233100302212100200010123330003100100111311123130222300 03
XX:Z:TGGGCAAGTCCTTCCGGTCTATGGGCCTCAGTTTCCCCAACTATAAAATGGGTTTGTGCACAGCATTCTCGGGGC
575_1420_134_F3	0	Chr13	74413135	250	75M	*	0	0
ATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCAG	*	Z0:i:25149	Z1:i:25149	CM:i:0	NM:i:3
AS:i:641	CS:Z:T33203222031321113111001001300033321010002001212203001332210302020212013 0012
XX:Z:ATCCGAGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCAG
630_1342_868_F3	0	Chr13	74413140	255	67M	*	0	0
AGAATGCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCC	HHHHHHHHHHHHHHHHHHHHHHGGGGGA>8AHGHHHHFFFGHFFFGHEEEEGGDGEG>>>>AAAAAA
XA:i:2	MD:Z:4A8T53	NM:i:2
289_347_655_F3	0	Chr13	74413145	255	70M	*	0	0
GCTGTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCAGAGCCG	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGGHHHHHHEEEEHEDEDGAAAAGEEEEF<<<<AAAA
XA:i:1	MD:Z:8T61	NM:i:1
575_1268_1764_F3	16	Chr13	74413148	255	43M	*	0	0	GTGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAG
@BBD>BDDD>AAAA8ABAAAAA<BBBGGGEGG@BBBEEEEEHG	XA:i:1	MD:Z:5T37	NM:i:1
289_1069_1663_F3	0	Chr13	74413149	255	67M	*	0	0
TGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCAGAGCCGT	HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGGGGG;;;;FEEEDEAAAAGBBBBE>>>>@
XA:i:1	MD:Z:4T62	NM:i:1
341_1280_573_F3	16	Chr13	74413149	255	62M	*	0	0
TGCACAAACCCATTTTATAGTTGGGGAAACTGAGGCCCATAGACCGGAAGGACTTGCCCAGA	@@@@@EEDDDEDDDDGFGFGHGGGGHHHHHHGGGGHEGGGHFFFFHHHHHHGHHHHGGGGHH	XA:i:1
MD:Z:4T57	NM:i:1
```

# Materials

- SAM/BAM files
- Linux OS
- Python Packages:
  - -Plotly
  - -Pysam
  - -Docopt
  - -BioPython

# Visualization in comparison to reference genome

- **Color codes to reference genome**
- **Reads Annotations**
- **Color codes to reads indicating direction (reverse complimented or forward)**
- **Read Coverage bar graph**
- **Paired End**
- **Visualization Start & End Regions**
- **Command Line UI**

# **Features:** Color codes to reference genome

**Having the capability to read the reference genome gene by gene in color codes each color representing a different gene type:**

**A – in red**
**C – in yellow**
**G – green**
**T – in blue**

**Which aids in a more detailed and precise comparison of the sequence assembly to the reference genome.**

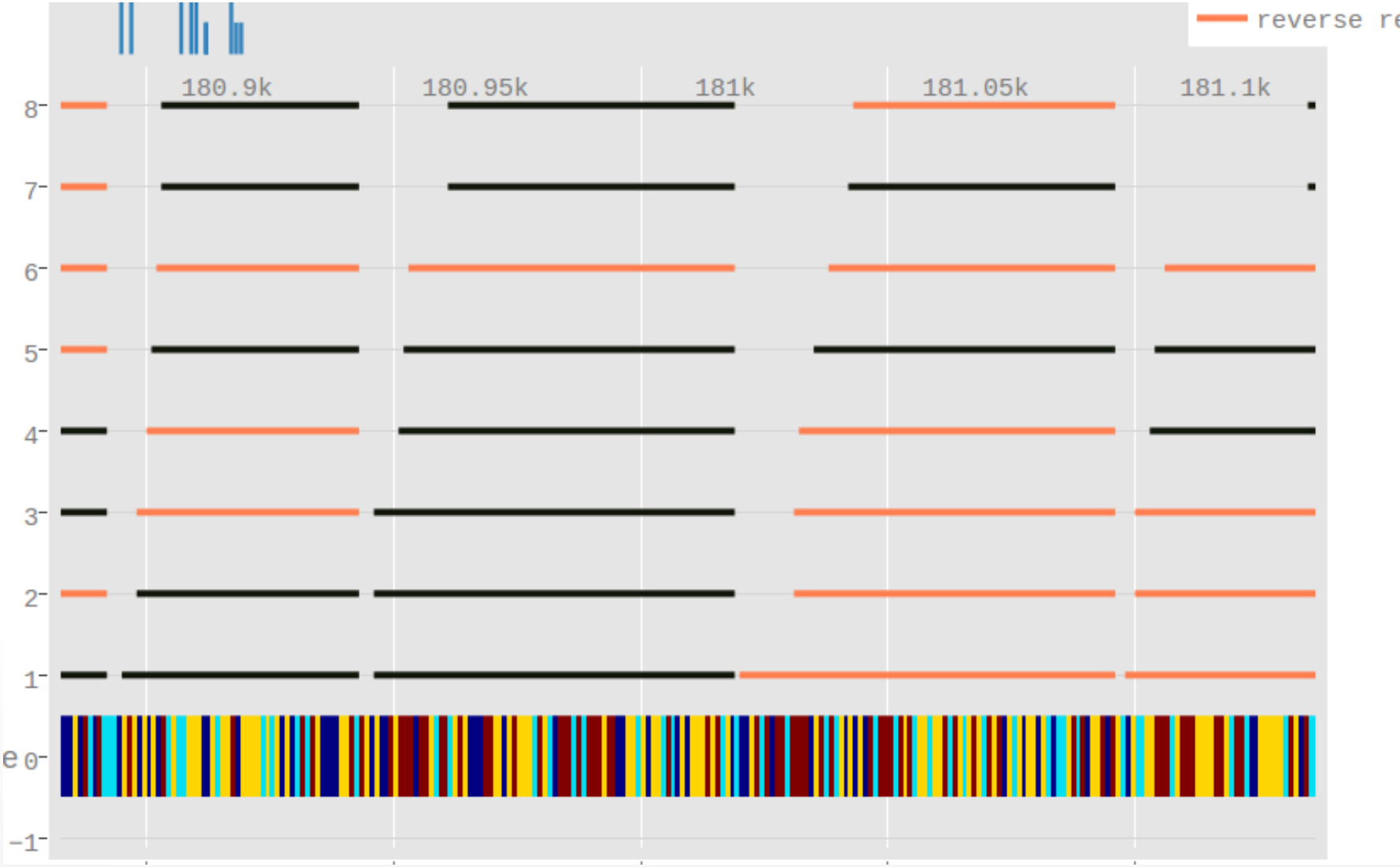
# **Features:**
## Reads Alignment Direction

**Reverse or Forward complimentary strands in the sequence alignment:**

**- An inversion in a sequence is a section of the DNA that is reversed in the subject genome compared to the reference genome.**

**- An inversion in paired-end reads are variant reads from the reference genome**

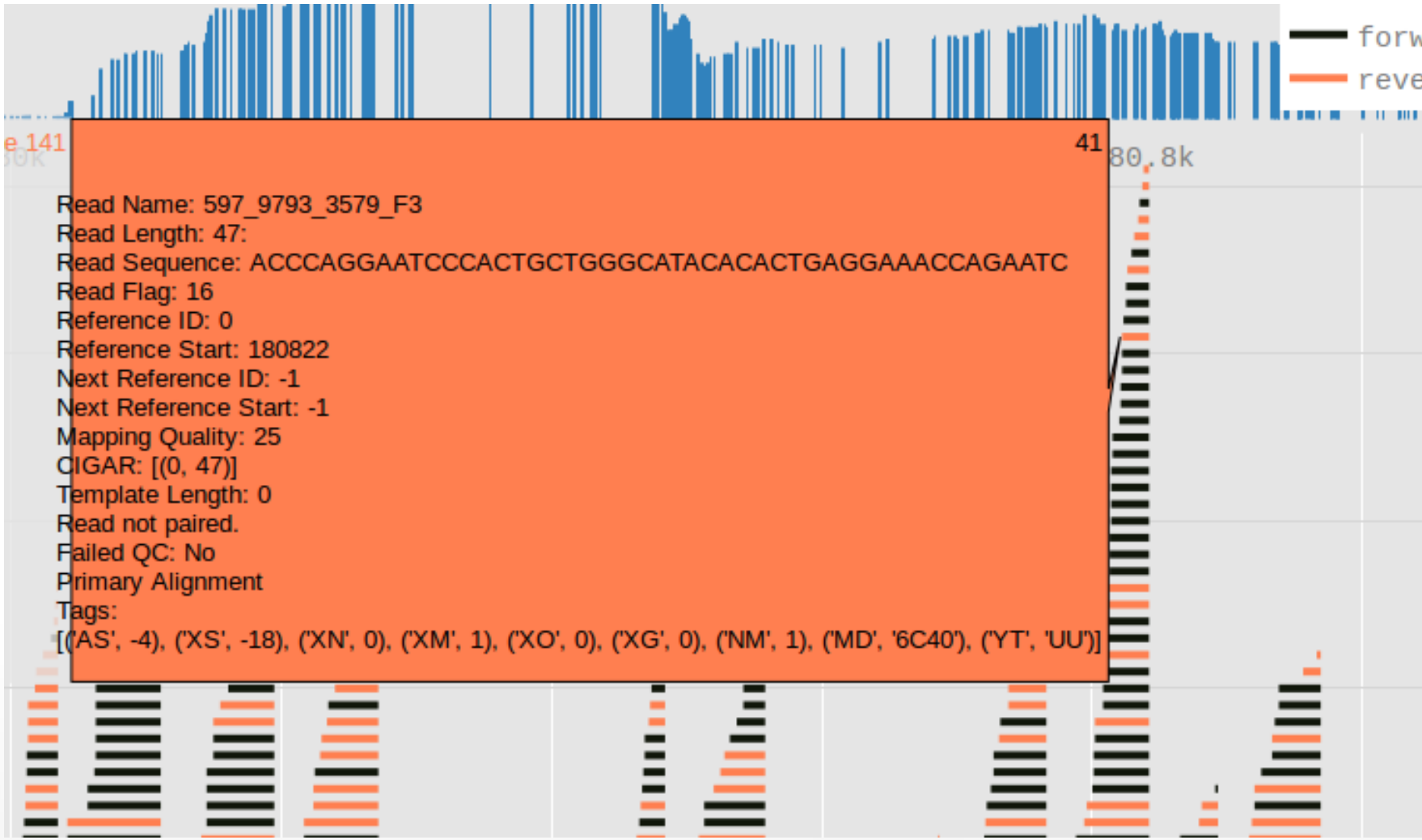
# **Results:** Color Codes



# **Features:**
## Reads Annotations

**Acquiring all the details of a single read by a single click or mouse positioning on the desired read from read position and sequence to the reference genome details.**

# **Results:** Annotations



# **Features:**
## Coverage and Depth

**<u>Coverage</u>: the number of times a genome has been sequenced.**

**Coverage or depth of sequencing needed depends on the application of sequencing and varies such as:**

- **For detecting mutations, SNPs, and rearrangements**

- **For RNA sequencing**

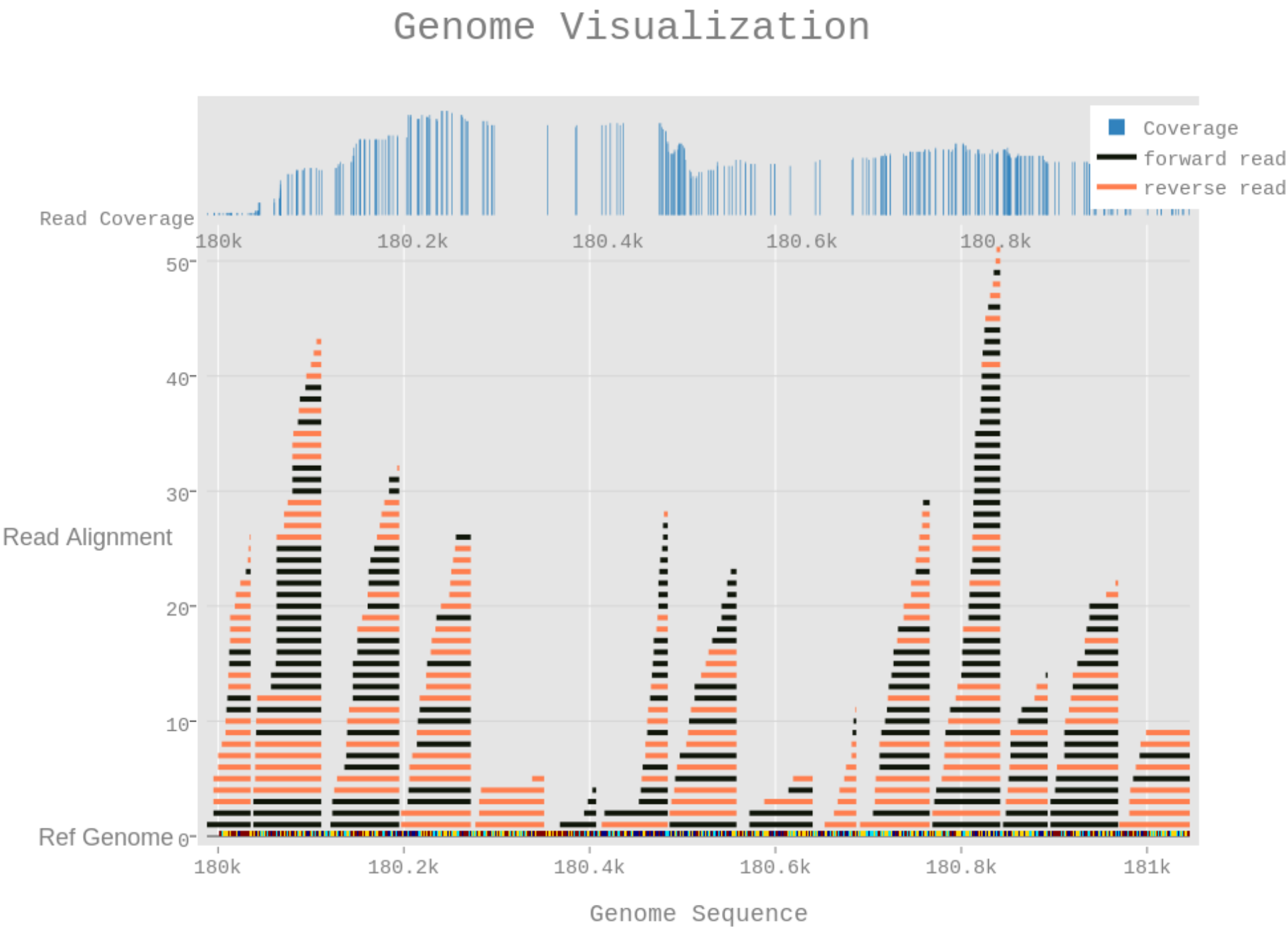
- **For ChIP-Seq**

# **Features:**
## Coverage and Depth

**When to sequence more?**

- **The effects seen are not statistically significant**
- **Events investigated are very rare**
- **Publication requires a higher level of coverage for a particular application**
- **Certain genomes may require more sequencing**

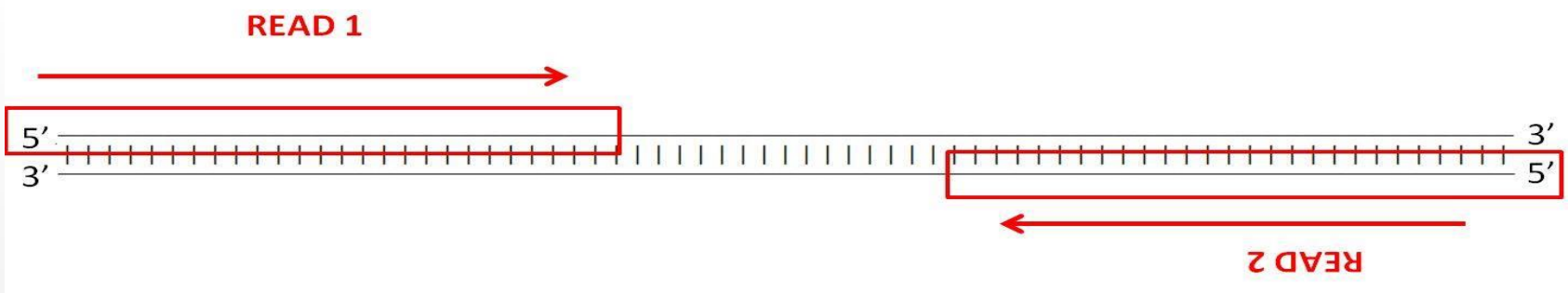
# Results


Genome Visualization
Coverage
forward read
reverse read
Read Coverage
180k
180.2k
180.4k
180.6k
180.8k
50
40
30
20
10
Read Alignment
Ref Genome
0
180k
180.2k
180.4k
180.6k
180.8k
181k
Genome Sequence

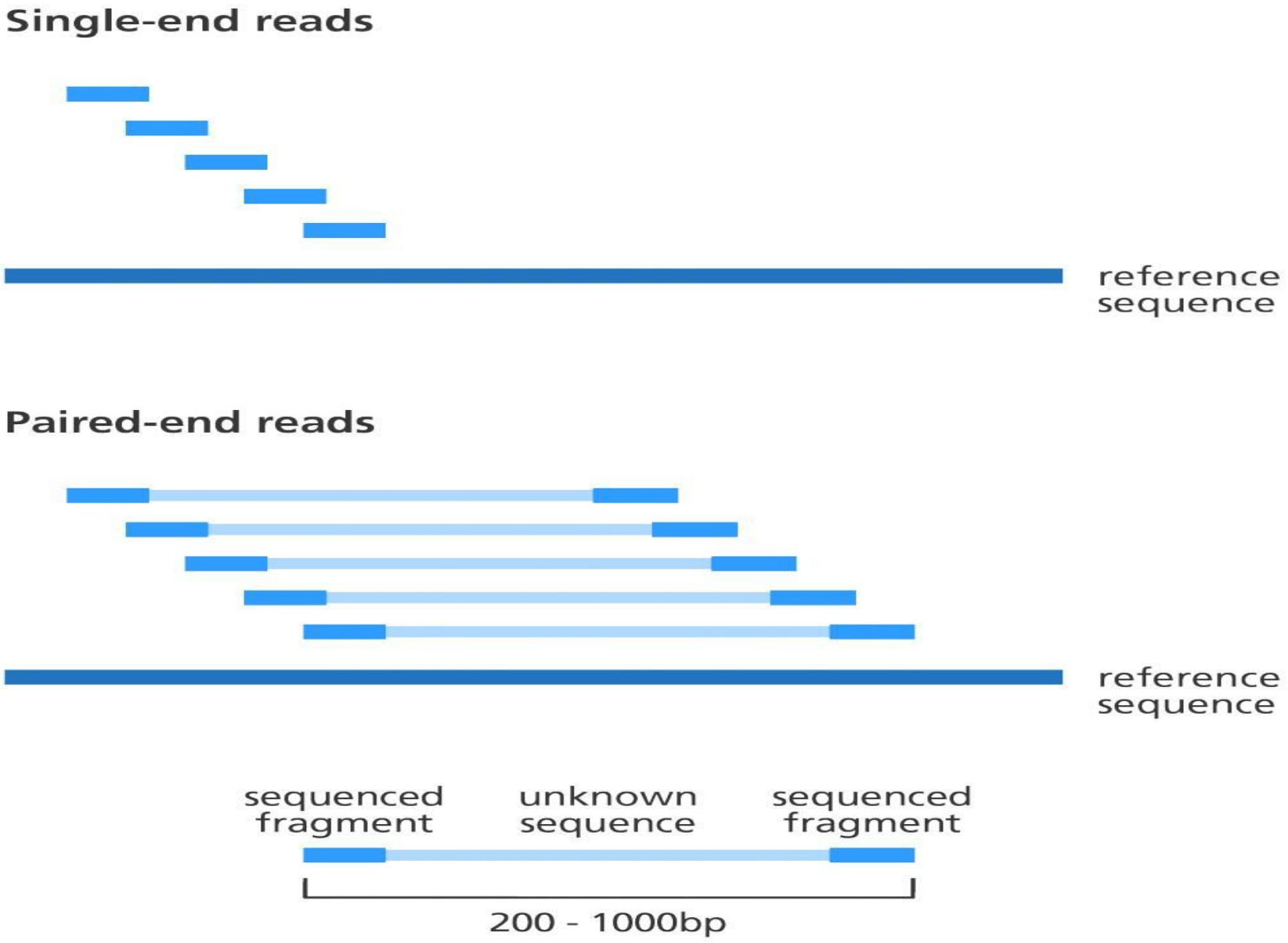
# **Features:**
## Paired-End

Paired-end sequencing allows users to sequence both ends of a fragment and generate high-quality, alignable sequence data.

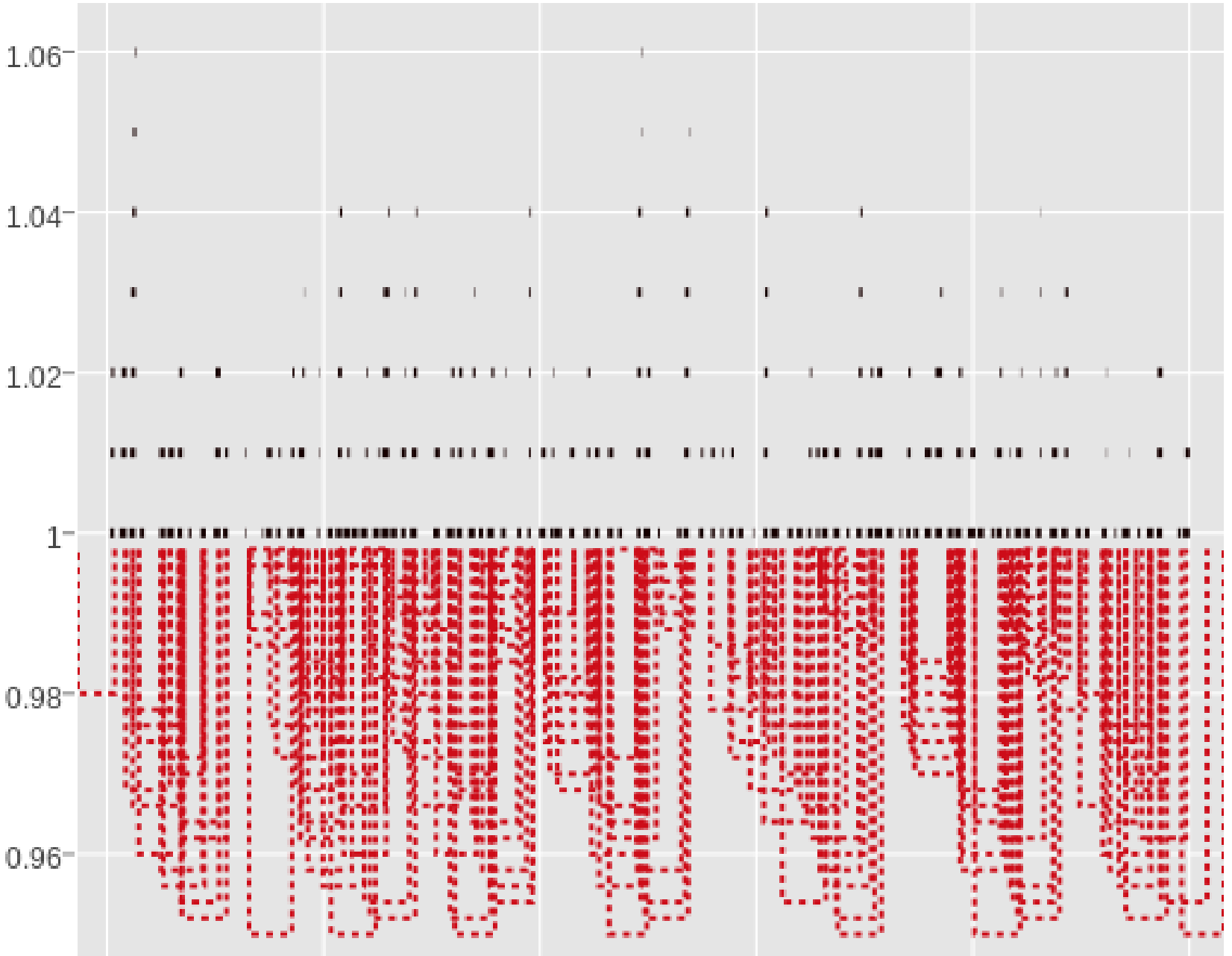
Paired-end sequencing facilitates detection of genomic rearrangements and repetitive sequence elements, as well as gene fusions and novel transcripts.



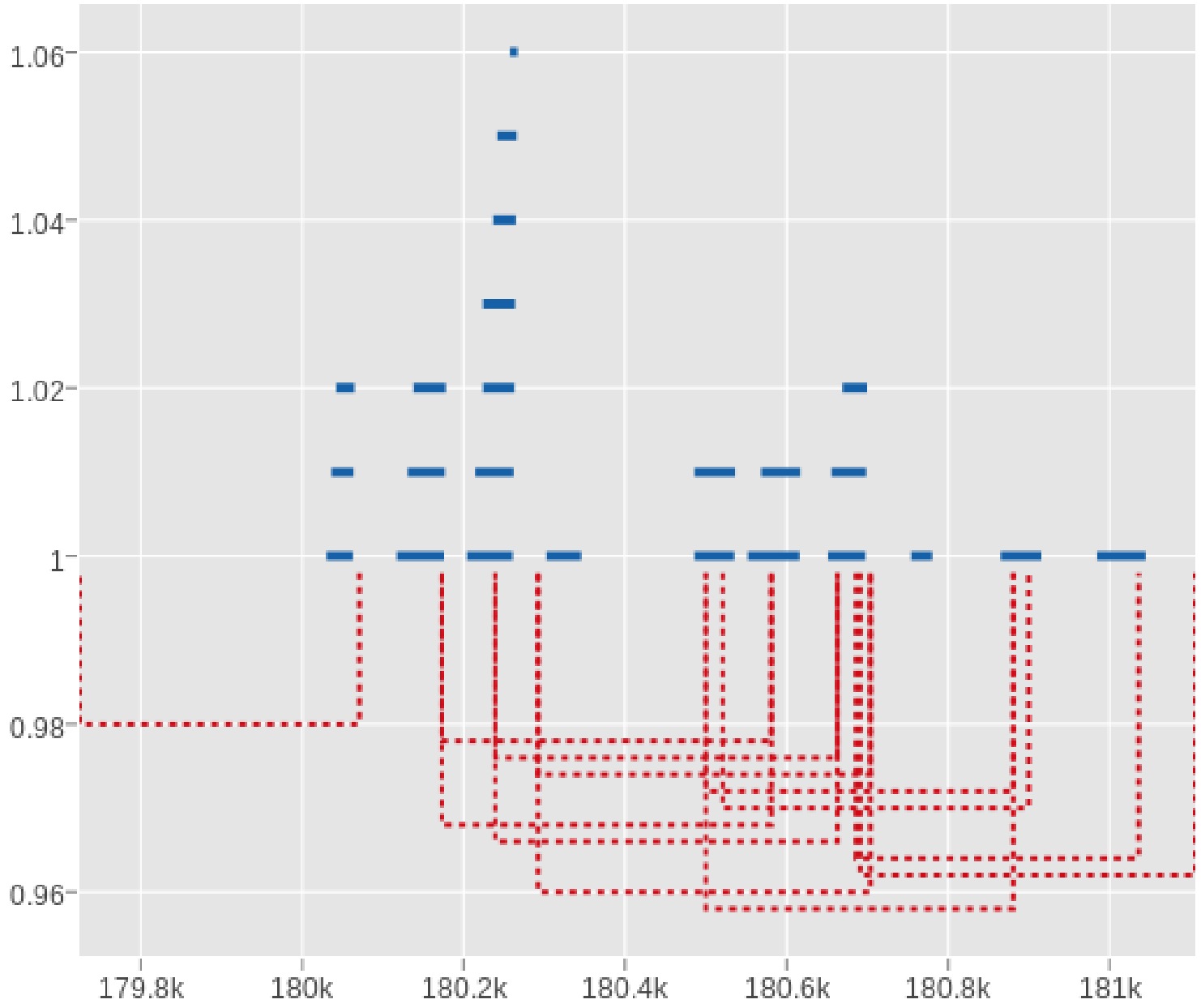
# **Features:**
# Paired-End



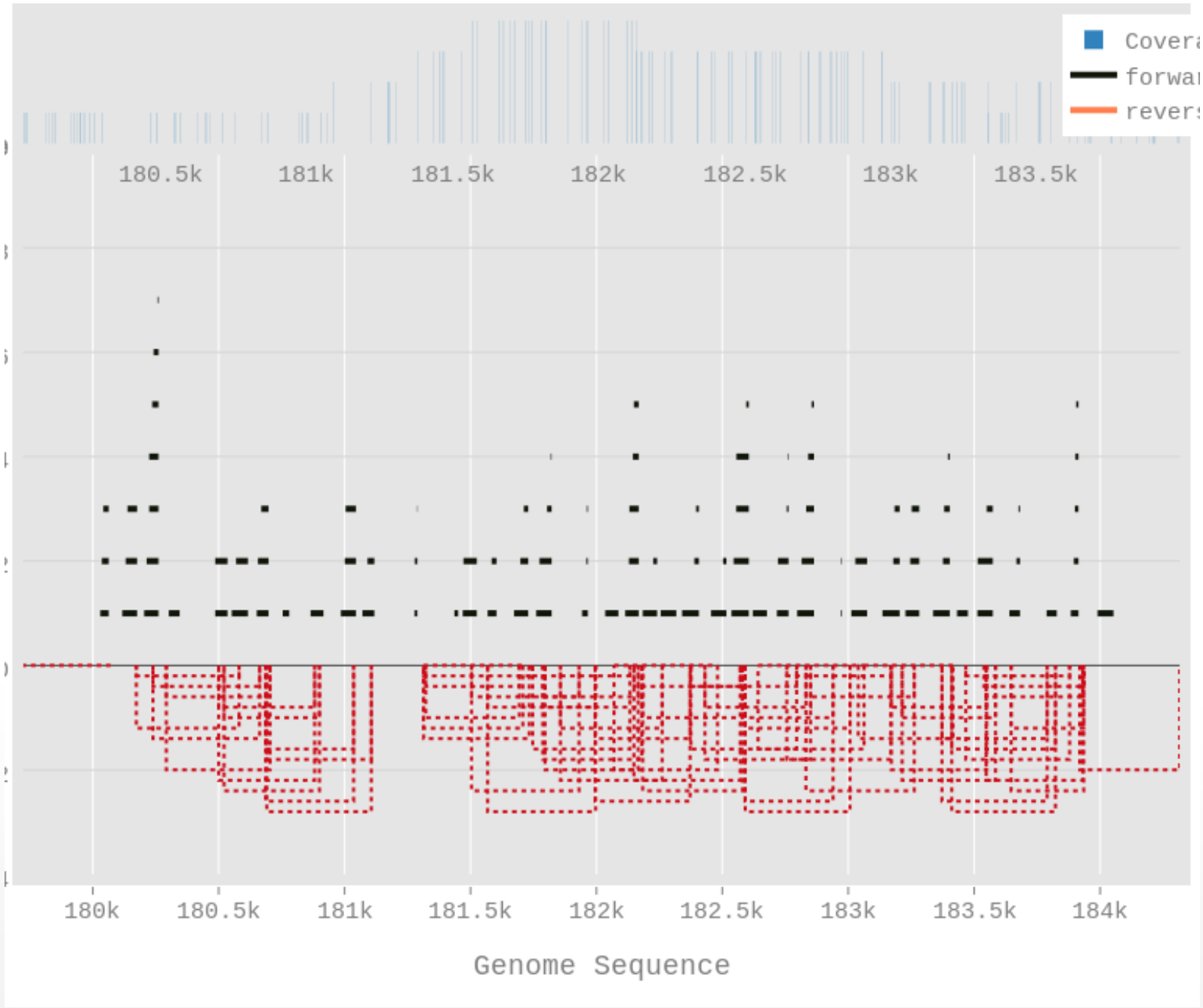
# Results: Paired End Lines



# Results: Paired End Lines



# Results: Paired End Lines



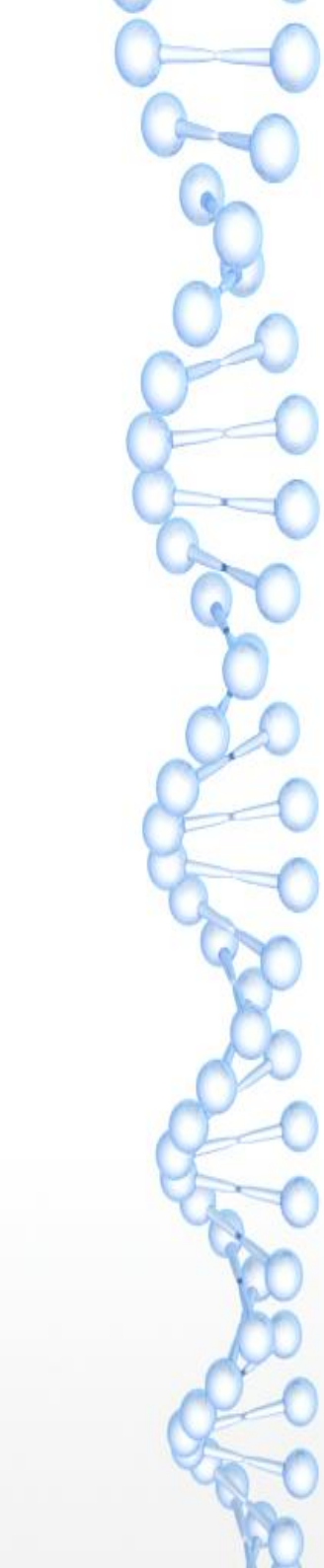
# Command Line UI

```
root@fatma:/home/fatma/Documents/PairedEnd/IGV# python test.py -help
Usage:
vst_main.py (-view FILE) [(-ref NAME --start VALUE --end VALUE )]  [-reff FASTAFILE] [--mp]
vst_main.py (-h | --help)

Visualizes or plots reads in SAM/BAM files to aid in the analysis of data through different
plotting manners with features enhancing analyzing effects


Arguments:
  FILE      input file to be visualized [BAM format]
  VALUE     start and end regions' values
  <path>      directory at which to execute the plots
  FASTAFILE   input fasta refeerence genome


Options:
  -view FILE                Imports and Views BAM files
  -ref  NAME                Name of the reference chromosome
  -s,--start VALUE          Determines specific starting region in file
  -e, --end VALUE           Determines the ending region in file (optional)[default: end]
  --mp                      Displays matepairs among plots
  -h, --help                Shows help document and quit
  -reff FASTAFILE           Plots reference genome
```



# THANK YOU !