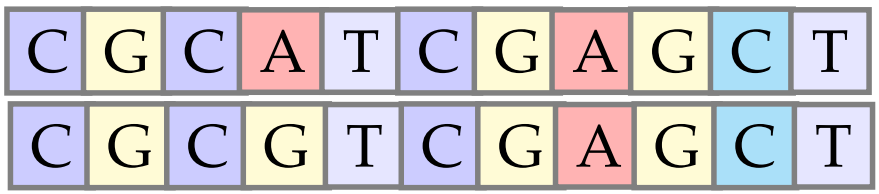
Pattern Recognition In Clinical Data

Saket Choudhary
Dual Degree Project

Guide: Prof. Santosh Noronha



October 30, 2013

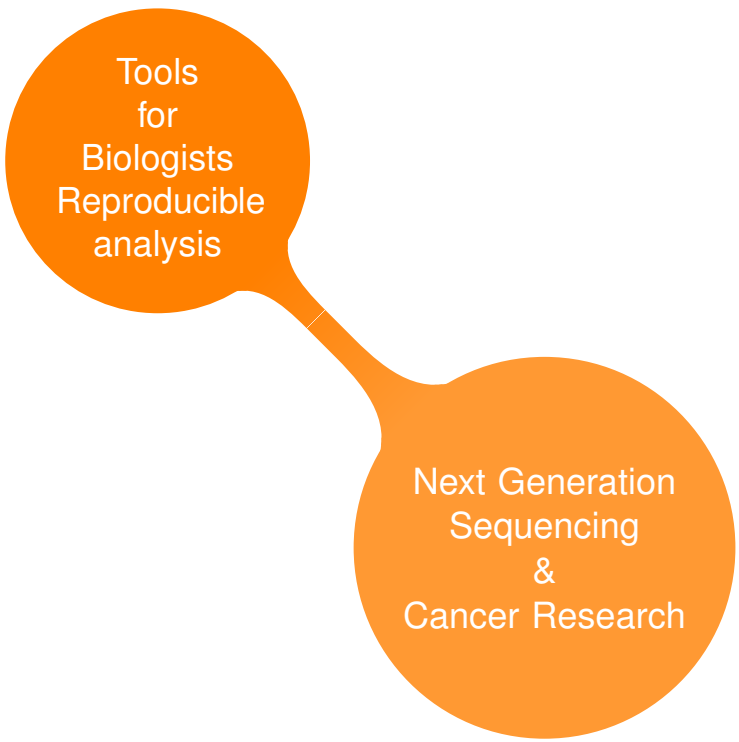
Introduction

OBJECTIVE

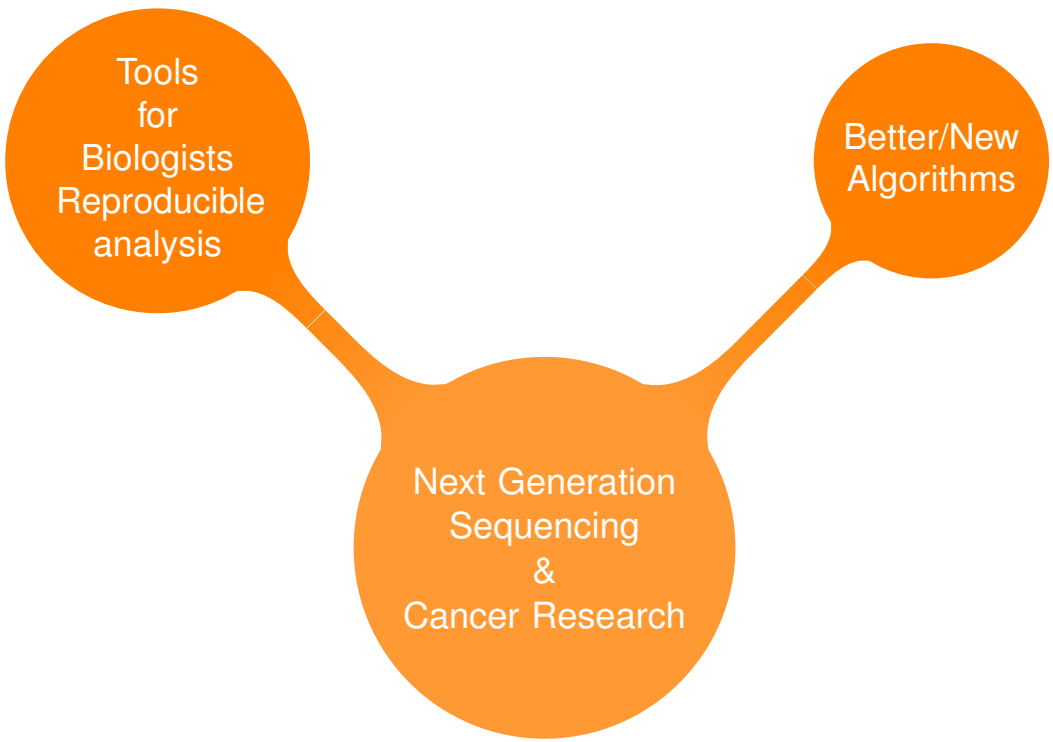
OBJECTIVE

Next Generation Sequencing & Cancer Research

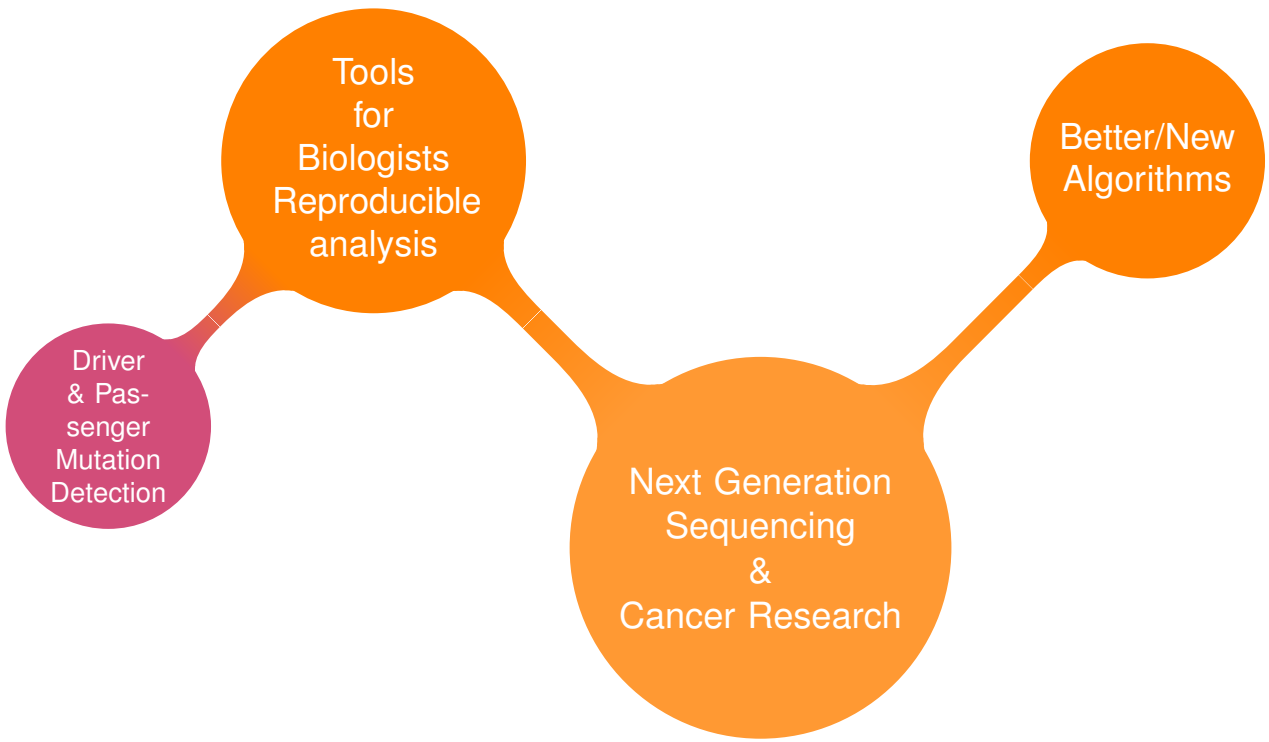
OBJECTIVE


Tools
for
Biologists
Reproducible
analysis
Next Generation
Sequencing
&
Cancer Research

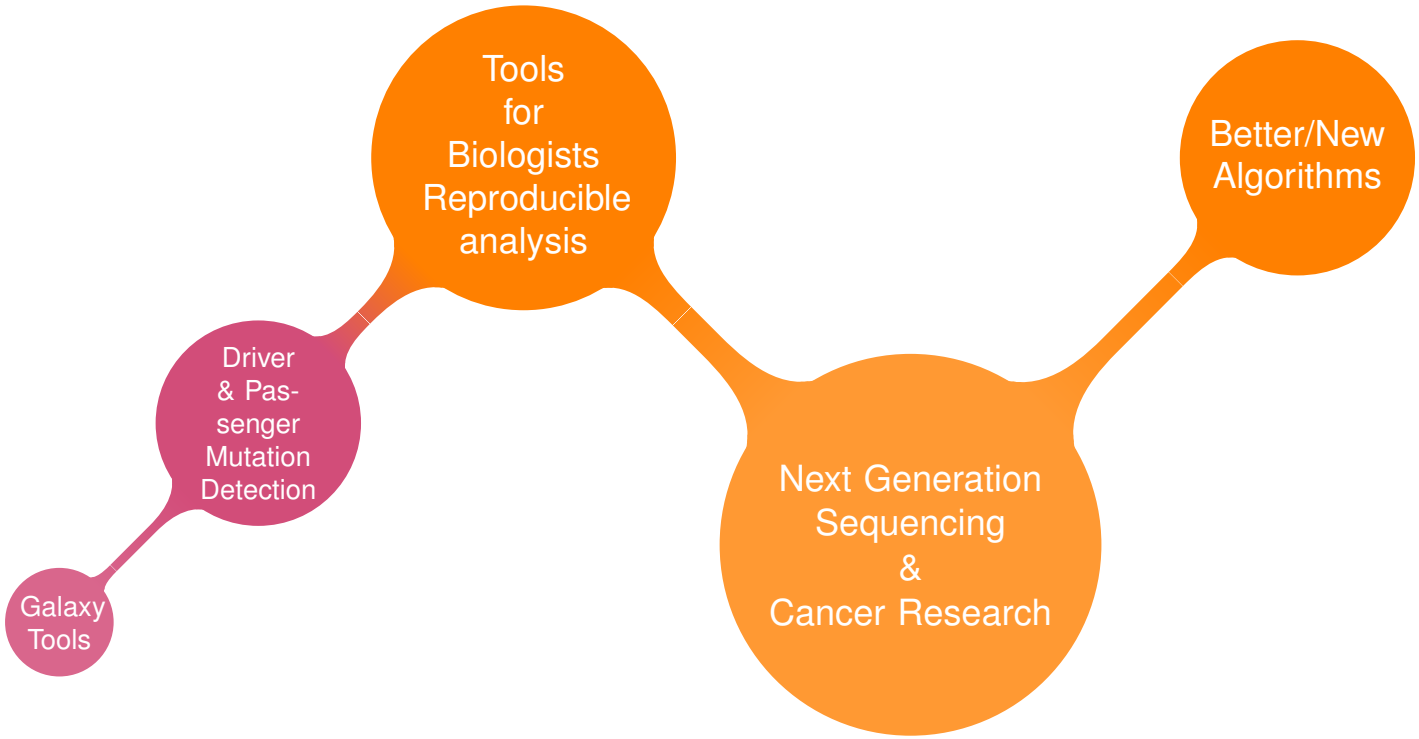
OBJECTIVE


Tools
for
Biologists
Reproducible
analysis
Better/New
Algorithms
Next Generation
Sequencing
&
Cancer Research

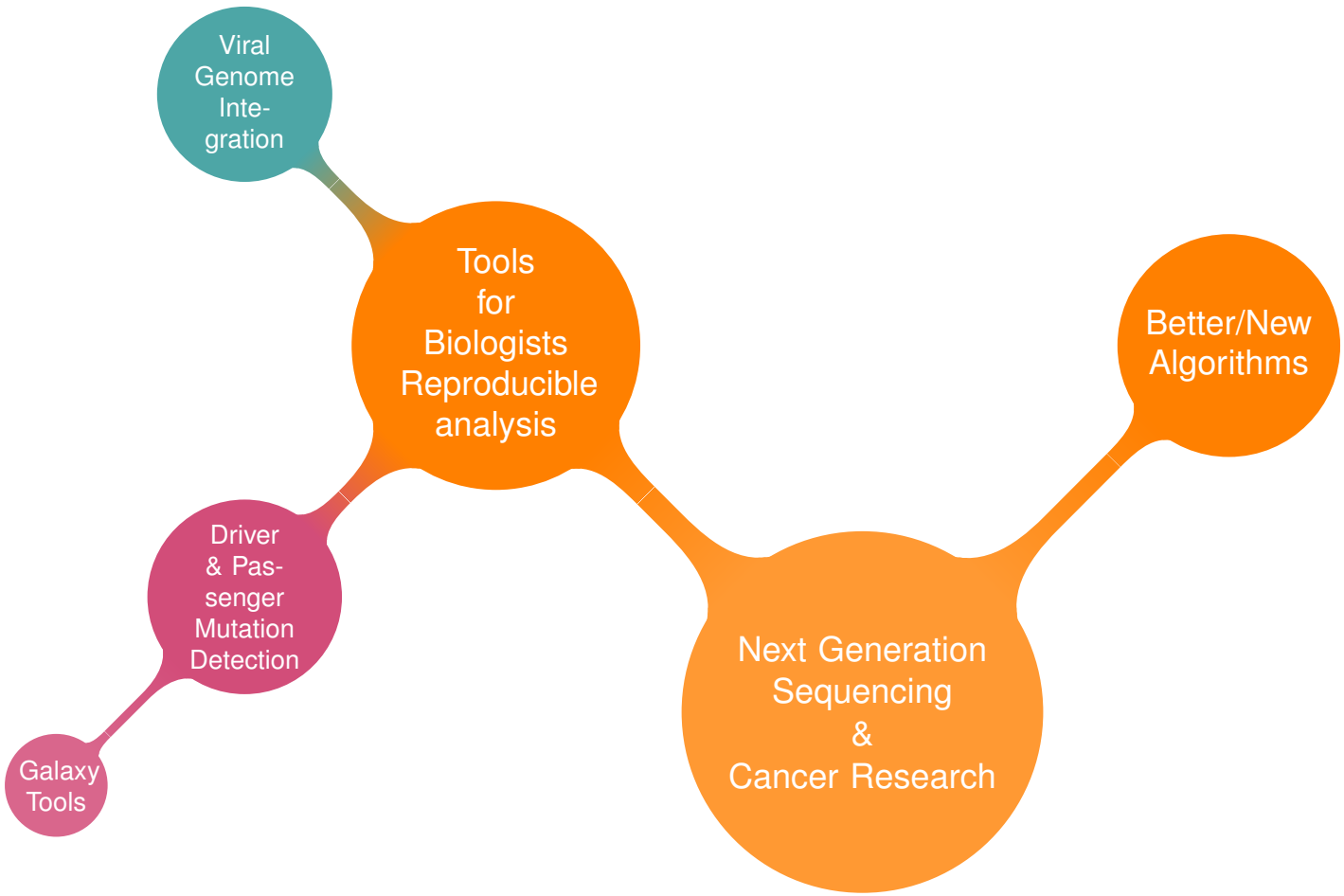
OBJECTIVE


Tools
for
Biologists
Reproducible
analysis
Better/New
Algorithms
Driver
& Pas-
senger
Mutation
Detection
Next Generation
Sequencing
&
Cancer Research

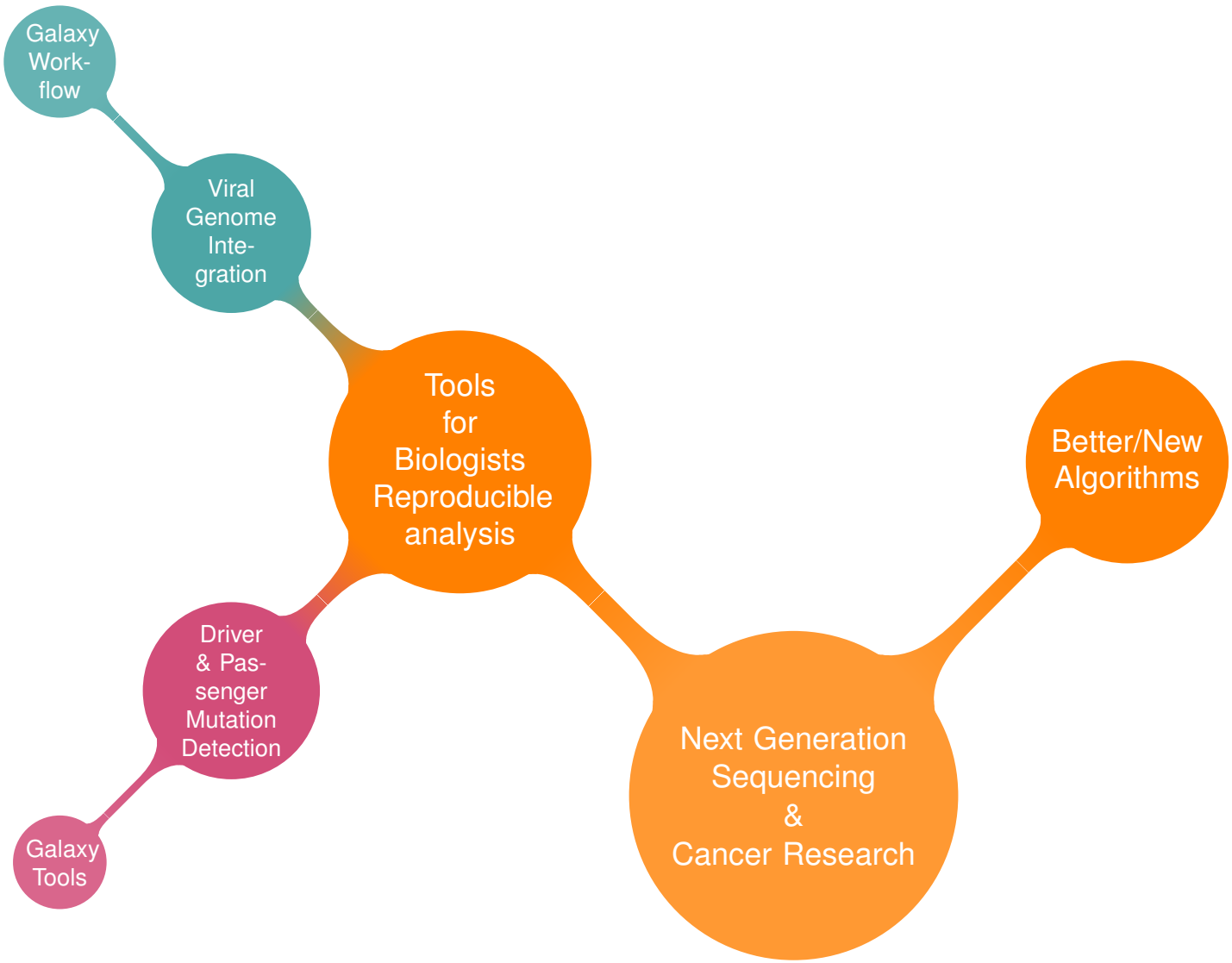
OBJECTIVE


Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Galaxy Tools

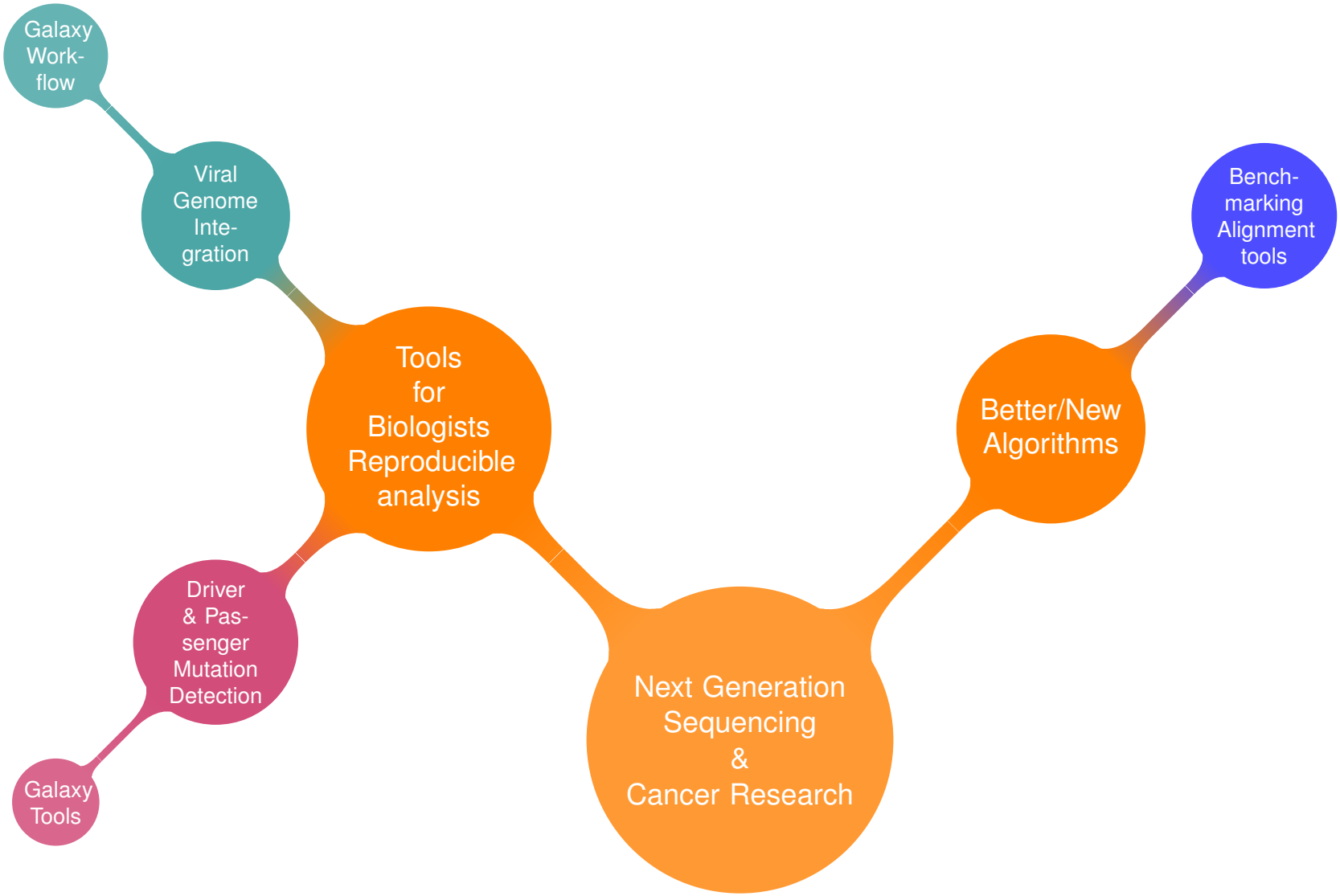
OBJECTIVE


Viral Genome Inte-gration
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Galaxy Tools

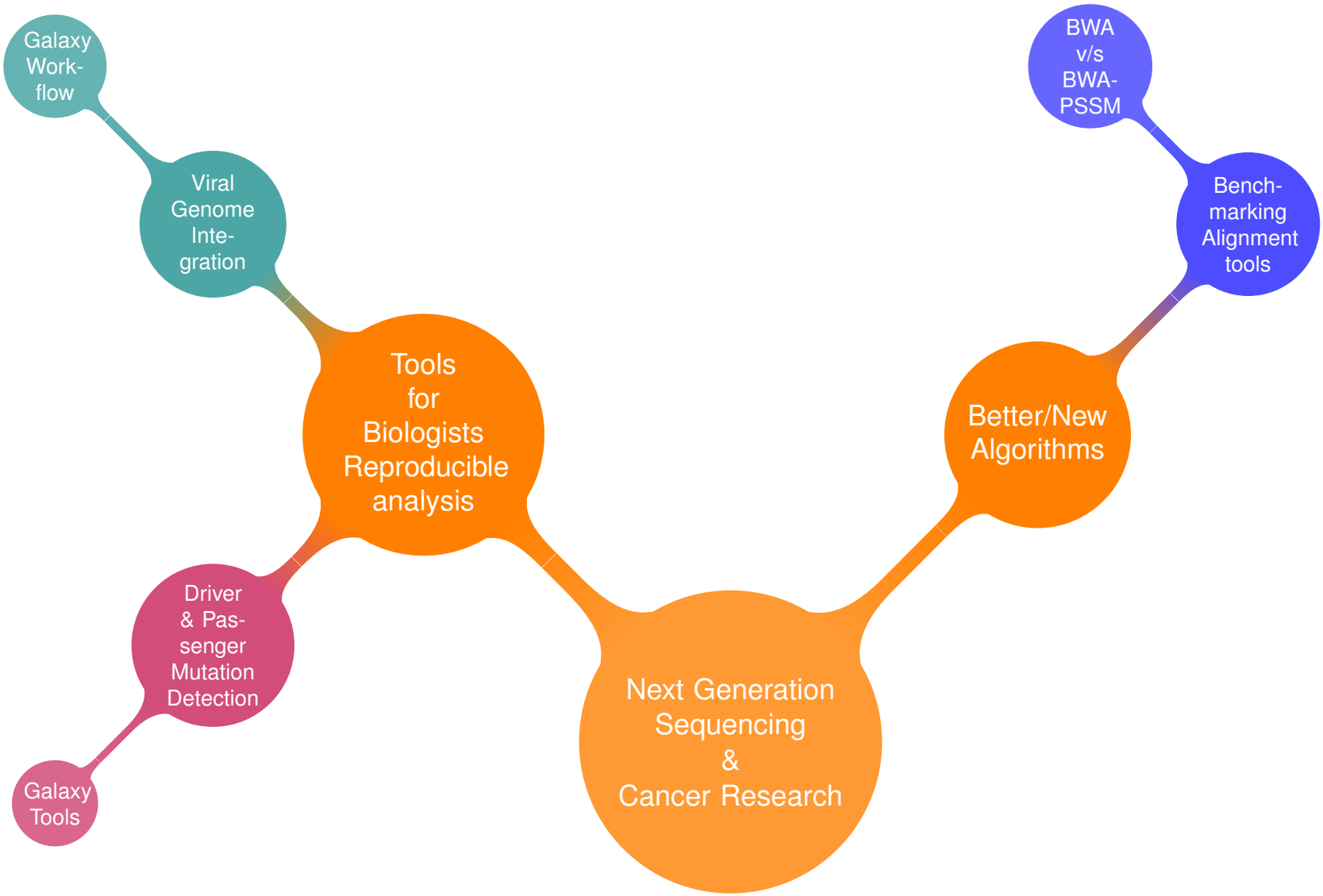
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Galaxy Tools

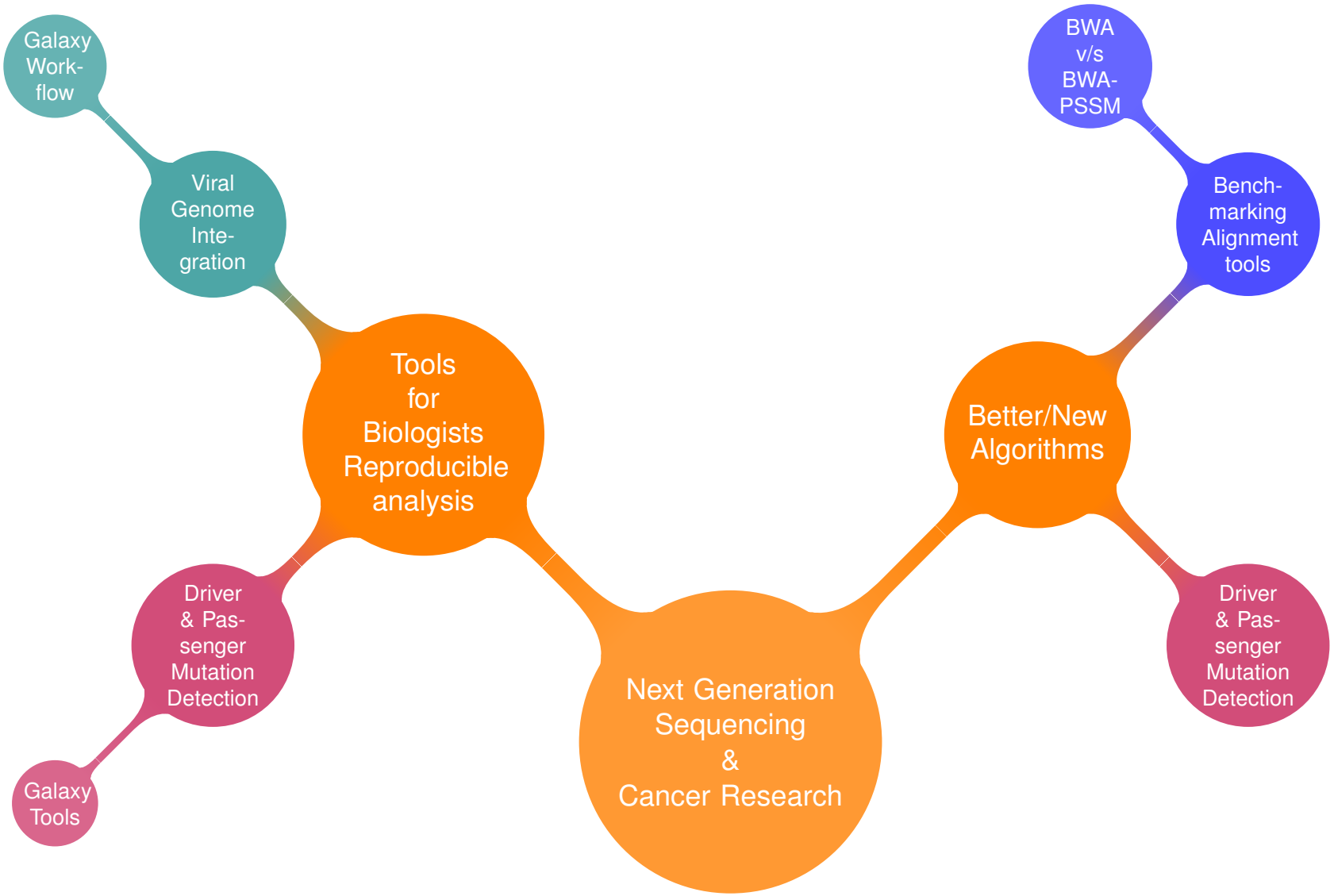
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
Tools for Biologists Reproducible analysis
Bench-marking Alignment tools
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Galaxy Tools
Next Generation Sequencing & Cancer Research

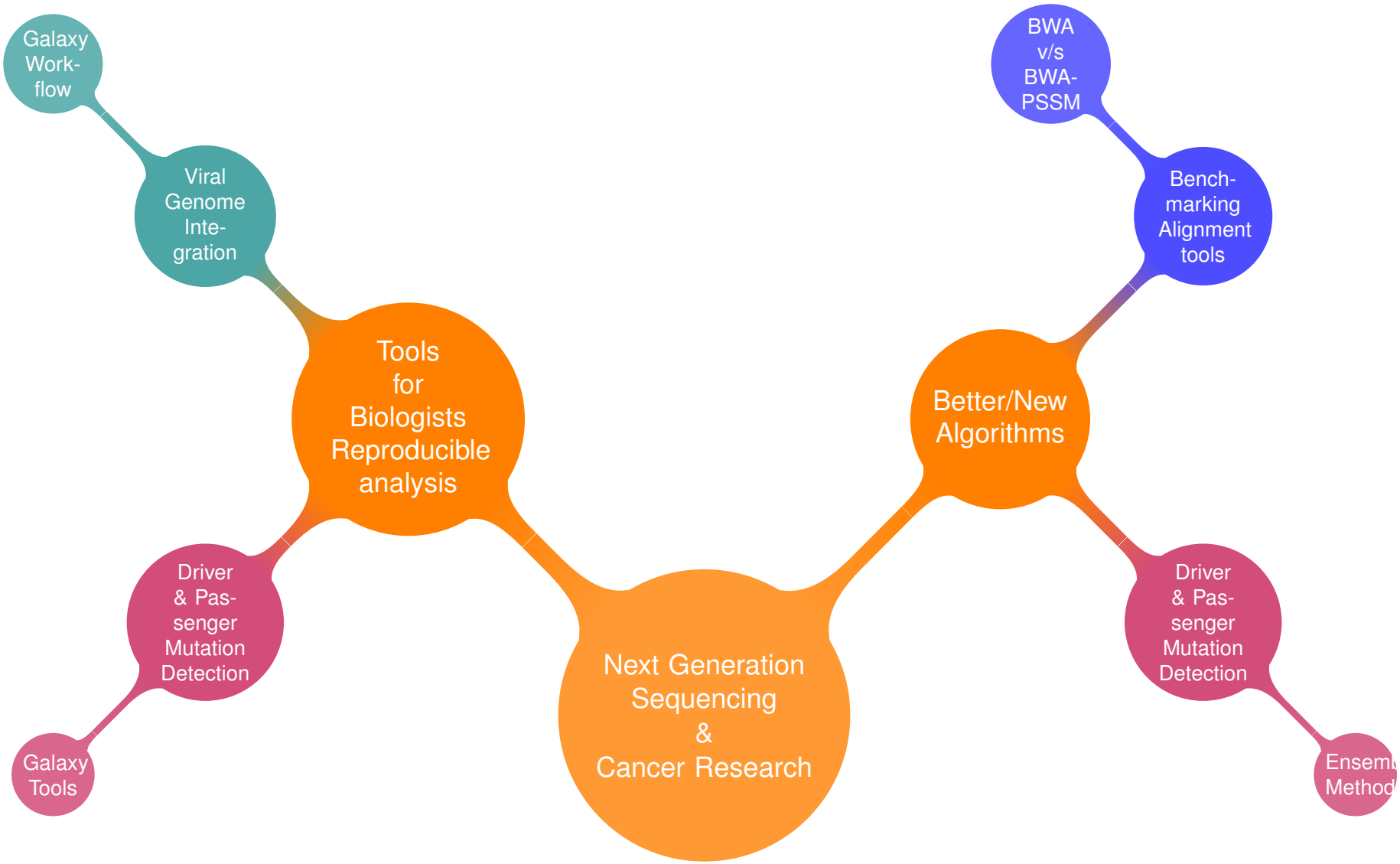
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
BWA v/s BWA-PSSM
Bench-marking Alignment tools
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Galaxy Tools
Next Generation Sequencing & Cancer Research

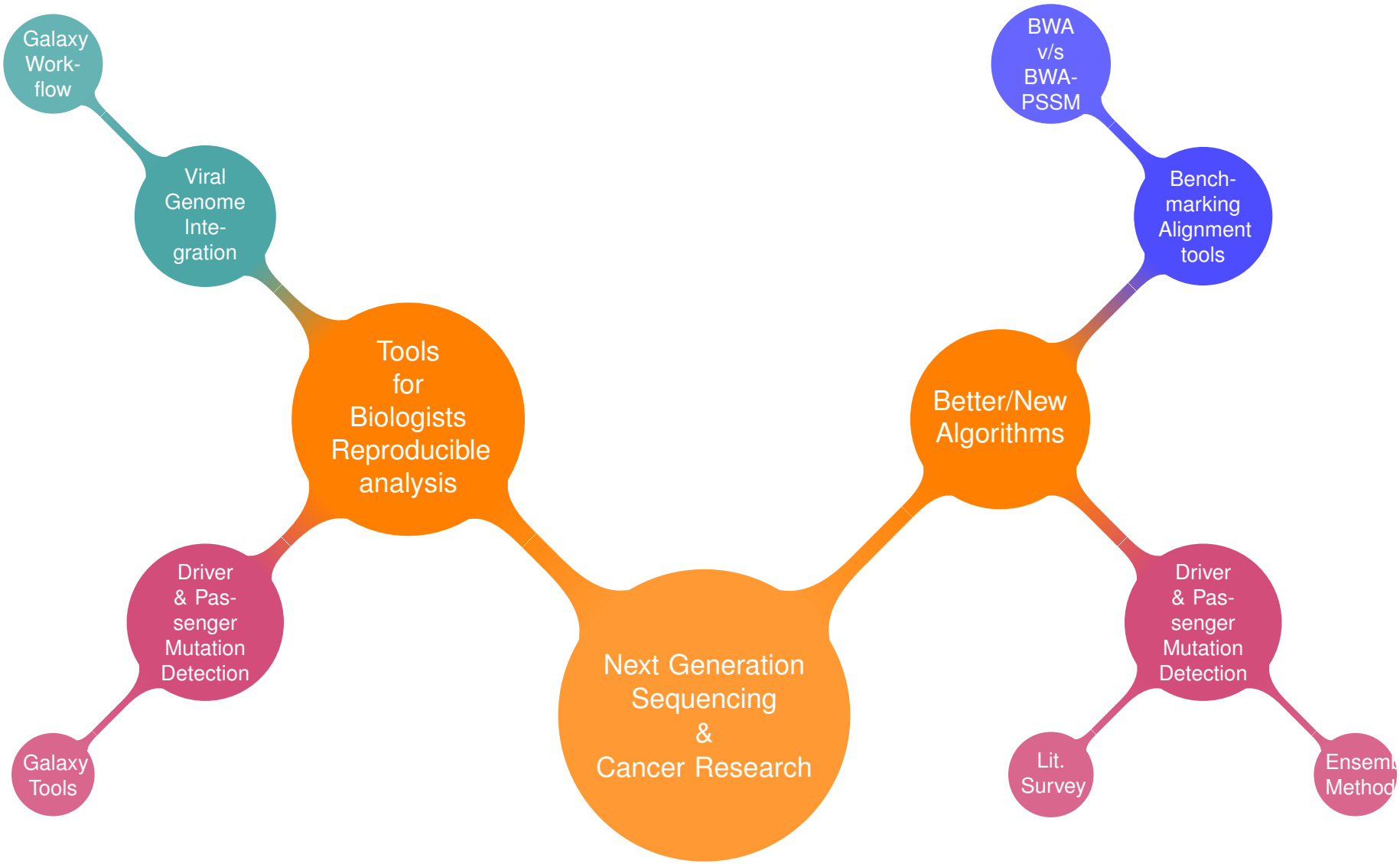
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
BWA v/s BWA-PSSM
Bench-marking Alignment tools
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Galaxy Tools

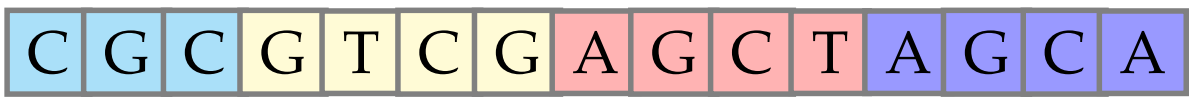
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
BWA v/s BWA-PSSM
Bench-marking Alignment tools
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Driver & Pas-senger Mutation Detection
Galaxy Tools
Ensemu
Method

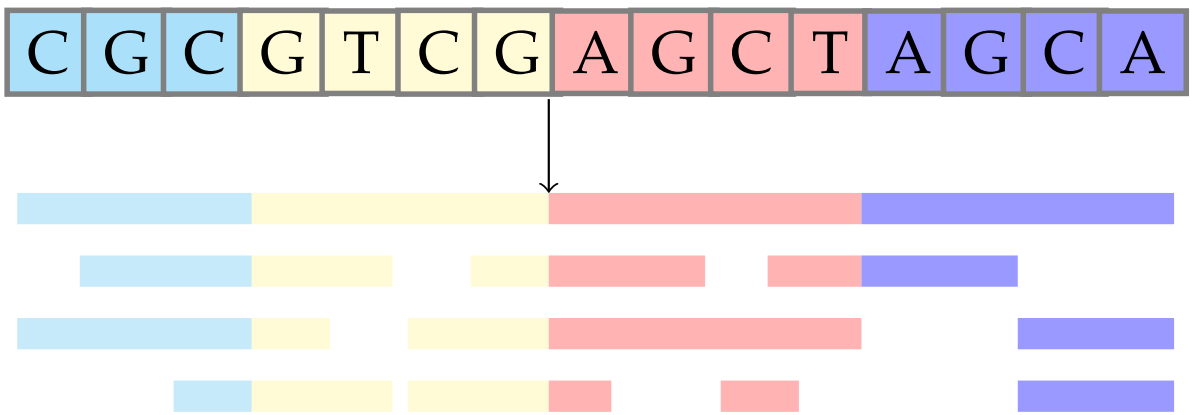
OBJECTIVE


Galaxy Work-flow
Viral Genome Inte-gration
BWA v/s BWA-PSSM
Bench-marking Alignment tools
Tools for Biologists Reproducible analysis
Better/New Algorithms
Driver & Pas-senger Mutation Detection
Next Generation Sequencing & Cancer Research
Driver & Pas-senger Mutation Detection
Galaxy Tools
Lit. Survey

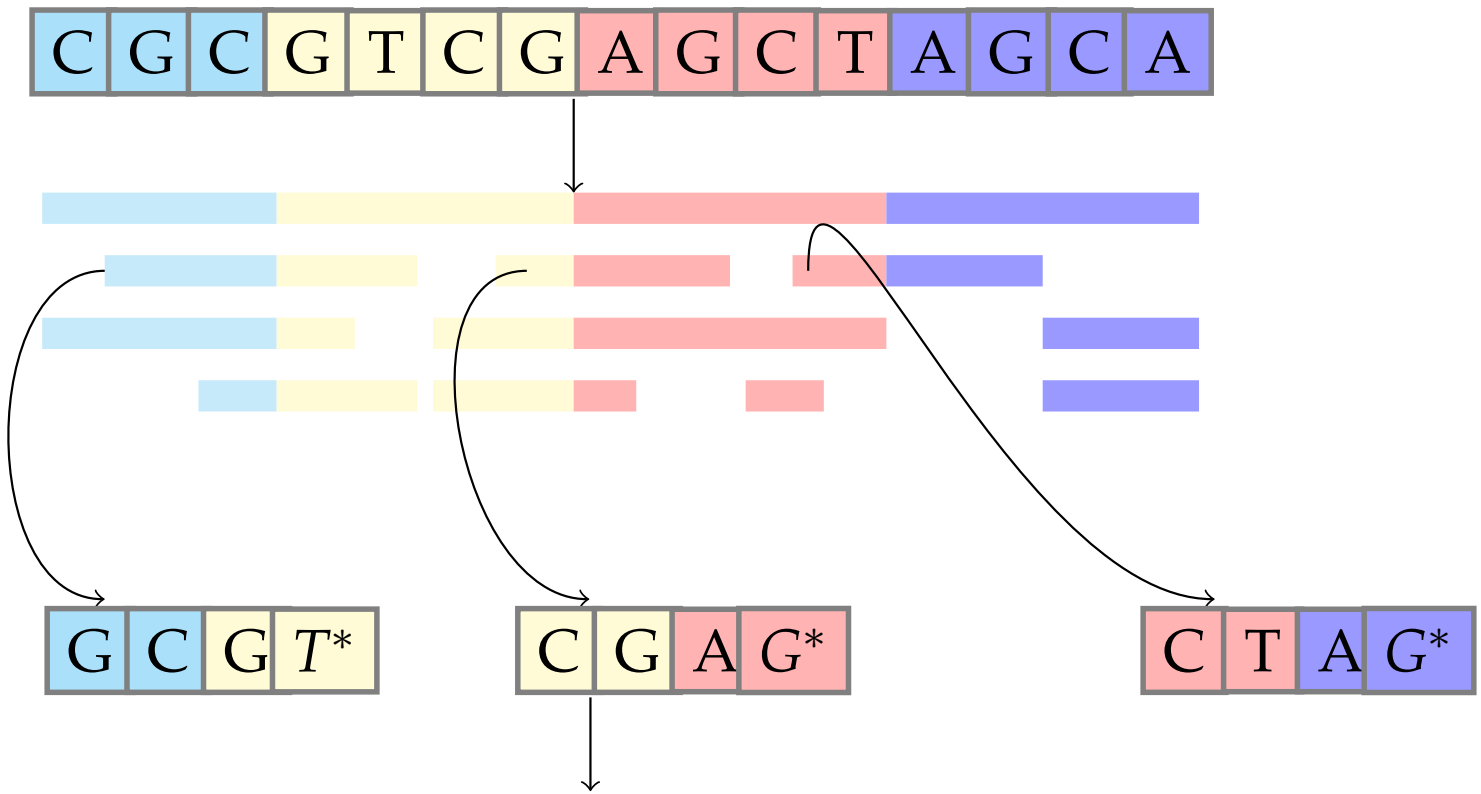
NEXT GENERATION SEQUENCING



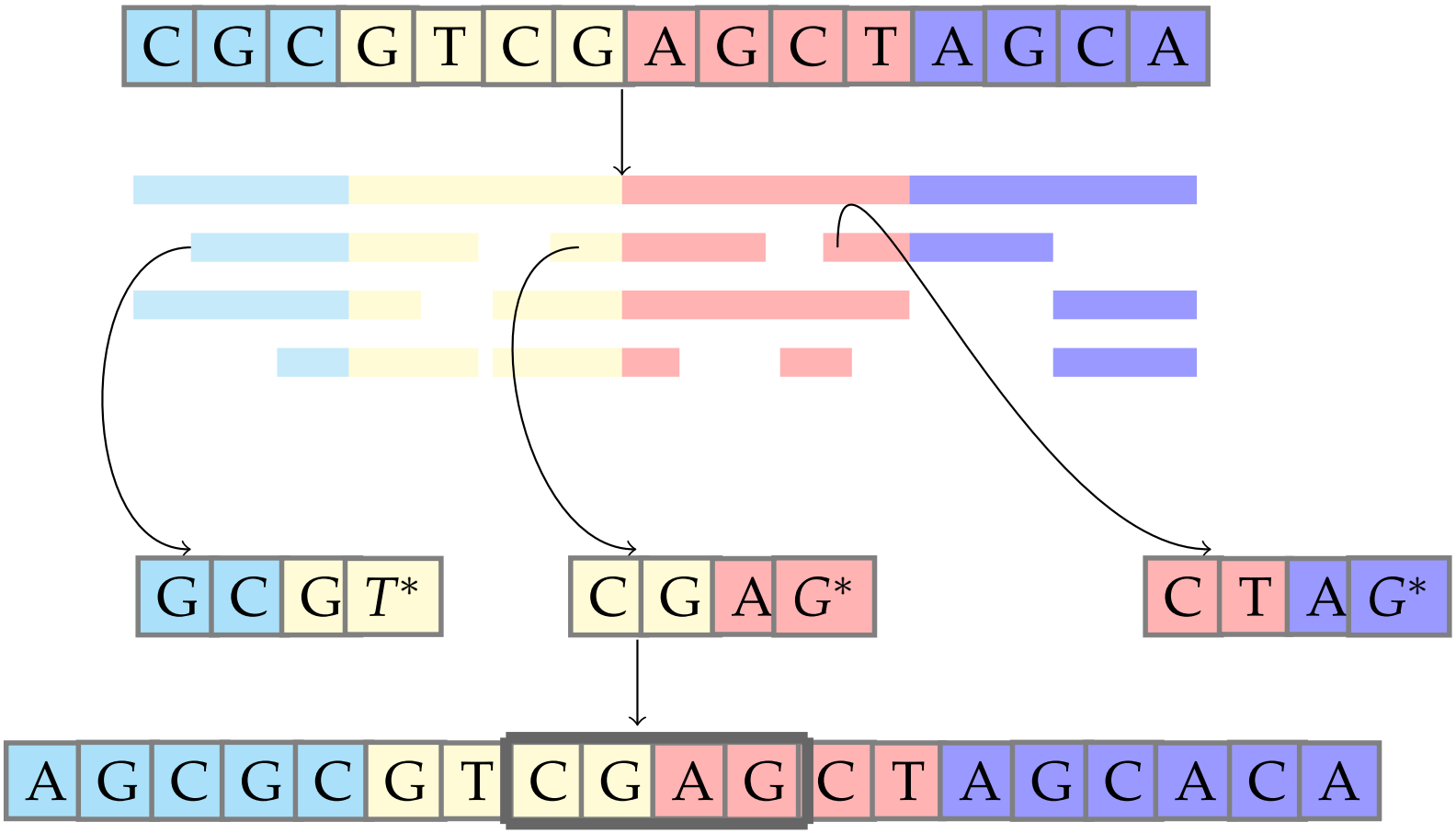
NEXT GENERATION SEQUENCING



NEXT GENERATION SEQUENCING



NEXT GENERATION SEQUENCING



NGS: WHY?

- **Molecular Approach**: Study of variations at the 'base' level
- **Low Cost**: 1000$ genome
- **Faster**: Quicker than traditional sequencing techniques

NGS: WHY?

- **Molecular Approach**: Study of variations at the 'base' level
- **Low Cost**: 1000$ genome
- **Faster**: Quicker than traditional sequencing techniques

NGS: WHY?

- **Molecular Approach**: Study of variations at the 'base' level
- **Low Cost**: 1000$ genome
- **Faster**: Quicker than traditional sequencing techniques

NGS: WHERE?

- Study variations, genotype-phenotype association
- Look for 'markers of diseases'
- Prognosis

NGS: WHERE?

- Study variations, genotype-phenotype association
- Look for 'markers of diseases'
- Prognosis

NGS: WHERE?

- Study variations, genotype-phenotype association
- Look for 'markers of diseases'
- Prognosis

NGS: MUTATIONS

- $3x10^9$ base pairs
- We are all 99.9% similar, at DNA level
- More than 2 million SNPs
- No particular pattern of SNPs
- If a certain mutation causes a change in an amino acid, it is referred to as non synonymous(nsSNV)

Drivers and Passengers I

Cancer is known to arise due to **mutations**
Not all mutations are equally important!

Somatic Mutations

Set of mutations *acquired* after zygote formation, over and above the **germline** mutations

Driver Mutations

Mutations that confer growth advantages to the cell, being selected positively in the tumor tissue

DRIVERS AND PASSENGERS

Drivers are **NOT** simply *loss of function* mutations, but more than that:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs

DRIVERS AND PASSENGERS

Drivers are **NOT** simply *loss of function* mutations, but more than that:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs

DRIVERS AND PASSENGERS

Drivers are **NOT** simply *loss of function* mutations, but more than that:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs

DRIVER MUTATIONS: WHY?

Identify **driver mutations** $\longrightarrow$ better therapeutic targets
But how does one zero down upon the exact set? $\longrightarrow$ experiments are too costly, probably infeasible for 2 million+ SNPs $\longrightarrow$ Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data analysis
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves $\longleftrightarrow$ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools

DRIVER MUTATIONS: WHY?

Identify **driver mutations** $\longrightarrow$ better therapeutic targets
But how does one zero down upon the exact set? $\longrightarrow$ experiments are too costly, probably infeasible for 2 million+ SNPs $\longrightarrow$ Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data analysis
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves $\longleftrightarrow$ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools

DRIVER MUTATIONS: WHY?

Identify **driver mutations** $\longrightarrow$ better therapeutic targets
But how does one zero down upon the exact set? $\longrightarrow$ experiments are too costly, probably infeasible for 2 million+ SNPs $\longrightarrow$ Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data analysis
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves $\longleftrightarrow$ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools

DRIVER MUTATIONS: WHY?

Identify **driver mutations** $\longrightarrow$ better therapeutic targets
But how does one zero down upon the exact set? $\longrightarrow$ experiments are too costly, probably infeasible for 2 million+ SNPs $\longrightarrow$ Leverage computational analysis

- Low cost of NGS comes with a heavier roadblock of data analysis
- Searching among 2 million+ SNPs is a non-trivial, and a computationally intensive problem
- Softwares have a low consensus ratio amongst them selves $\longleftrightarrow$ Defining a driver, computationally is non-trivial
- However there is no tool that allows one to visualise the results on an input across the cohort of tools

MACHINE LEARNING I

Two datasets:

- **Training**: *Labeled* dataset, containing a table of features with mutations labelled as "drivers/passengers"
- **Test**: 'Learning' from training dataset, test the prediction model

Table: Training Dataset

Chromosome	Position	Ref	Alt	Type
1	27822	A	G	Driver
1	27832	T	G	Driver
2	47842	G	C	Passenger
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Machine Learning II

Table: Test Dataset

Chromosome	Position	Ref	Alt	Type
1	27824	A	G	?
1	47832	T	G	?

MACHINE LEARNING: FEATURE SELECTION I

Machine Learning relies on a set of **features** for training
Redundant features should be avoided
CHASM [1] makes use of
$p(X_i)$ represents the probability of occurrence of an event X_i
Considering a series of events X_1, X_2, X_3...,X_n analogous 'series of packets' in communication theory , the information received at each step can be quantified on a log scale by:

$$\frac{1}{log_2(X_i)} = -log_2(p(X_i)) \tag{1}$$

The expected value of information from a series of events is called shannon entropy: $H(X)$:

$$H(X) = -\sum_i p(X_i)\log_2 p(X_i) \tag{2}$$

MACHINE LEARNING: FEATURE SELECTION II

Mutual Information between two random variables X, Y is defined as the amount of information gained about random variable X due to additional information gained from the second, Y:

$$I(X, Y) = H(X) - H(X|Y) \tag{3}$$

Here:
X: Class Label[Driver/Passenger]
Y: Predictive Feature
and hence $I(X, Y)$ represents how much information was gained about the class label Y from knowledge of a feature X
Simplifying :

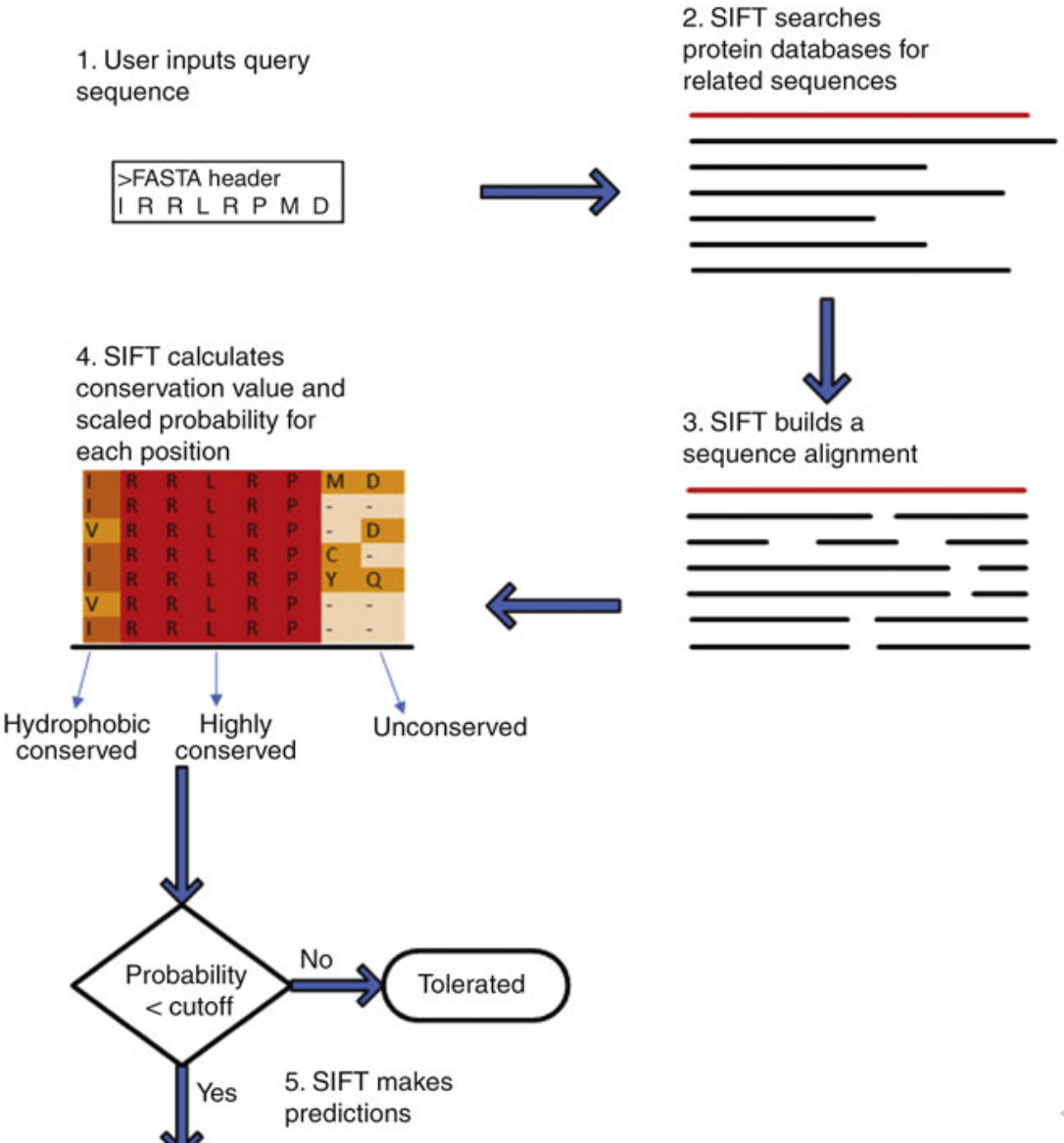
$$I(X, Y) = \sum p(x, y) log_2 \frac{p(x, y)}{p(x)p(y)} \tag{4}$$

FUNCTIONAL IMPACT I

- If a certain mutation confers an advantage to the cell in terms of replication rate, it is probably going to be selected while all those mutations that reduce its fitness have a higher chance of being eliminated from the population.
- Certain residues in a MSA of homologous sequences are more conserved than others. A highly conserved if mutated is possibly going to **cost a lot** since what had 'evolved' is disturbed!
- Scores can be assigned based on this "conservation" parameter.

FUNCTIONAL IMPACT II

Figure: SIFT [?] algorithm



Some of the common tools/algorithms used for driver mutation prediction:

- SIFT
- Polyphen
- Mutation Assesor
- TransFIC
- Condel

FRAMEWORK FOR COMPARING VARIOUS TOOLS I

- Different tools use different formats, give different outputs for similar input
- Running analysis on multiple tools $\longrightarrow$ keep shifting data formats
- Concordance?

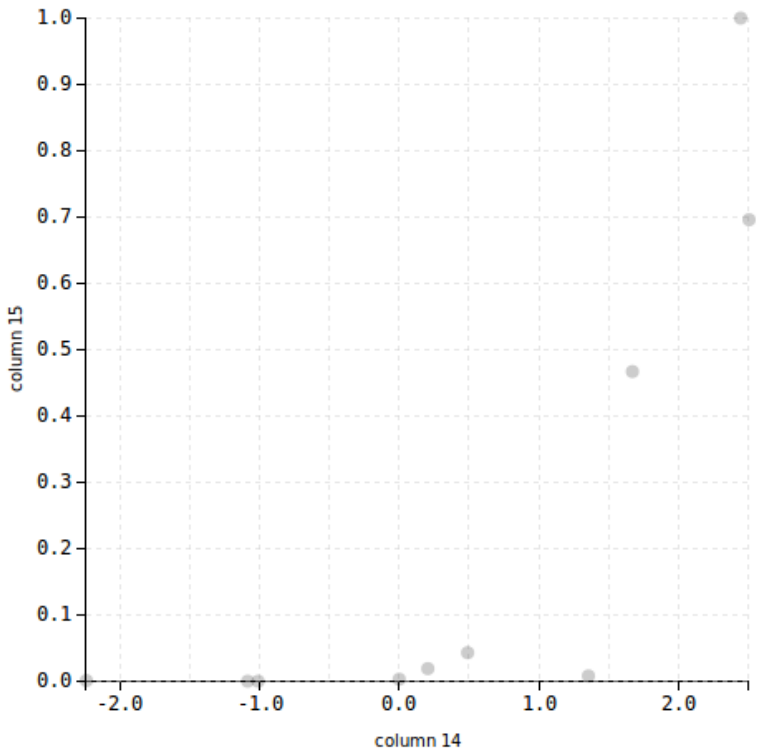
Polyphen2 Input

chr1:888659 T/C
chr1:1120431 G/A
chr1:1387764 G/A
chr1:1421991 G/A
chr1:1599812 C/T
chr1:1888193 C/A
chr1:1900186 T/C

FRAMEWORK FOR COMPARING VARIOUS TOOLS II

SIFT Input
1,888659,T,C
1,1120431,G,A
1,1387764,G,A
1,1421991,G,A
1,1599812,C,T
1,1888193,C,A
1,1900186,T,C

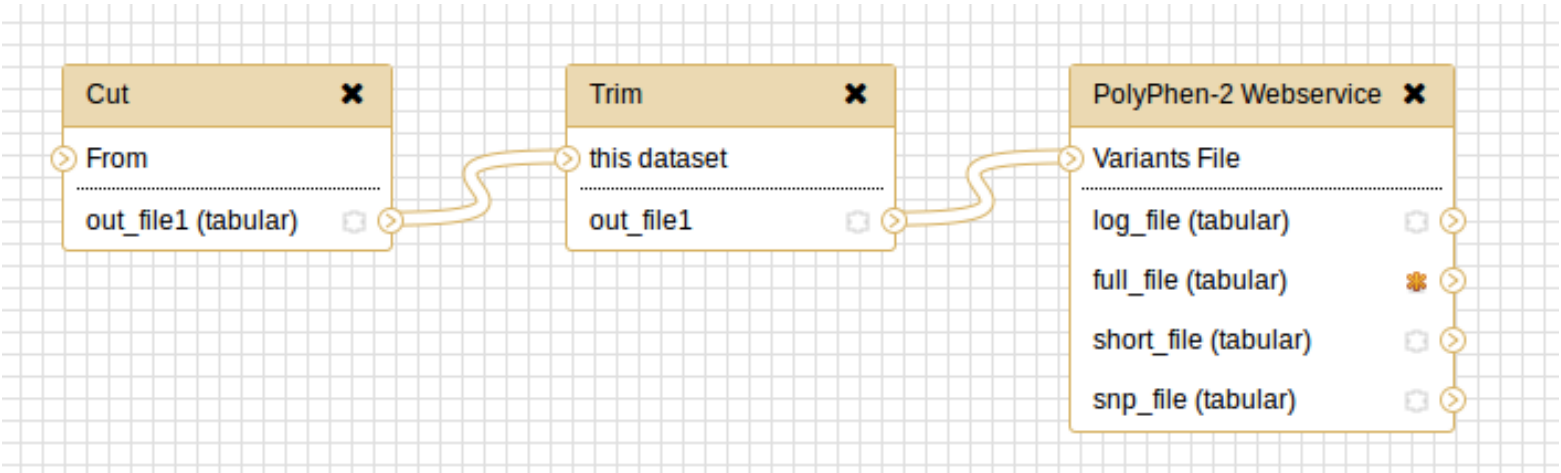
DRIVER MUTATIONS: TOOLS DON'T AGREE



X Axis: Condel Score Y Axis: MA Score

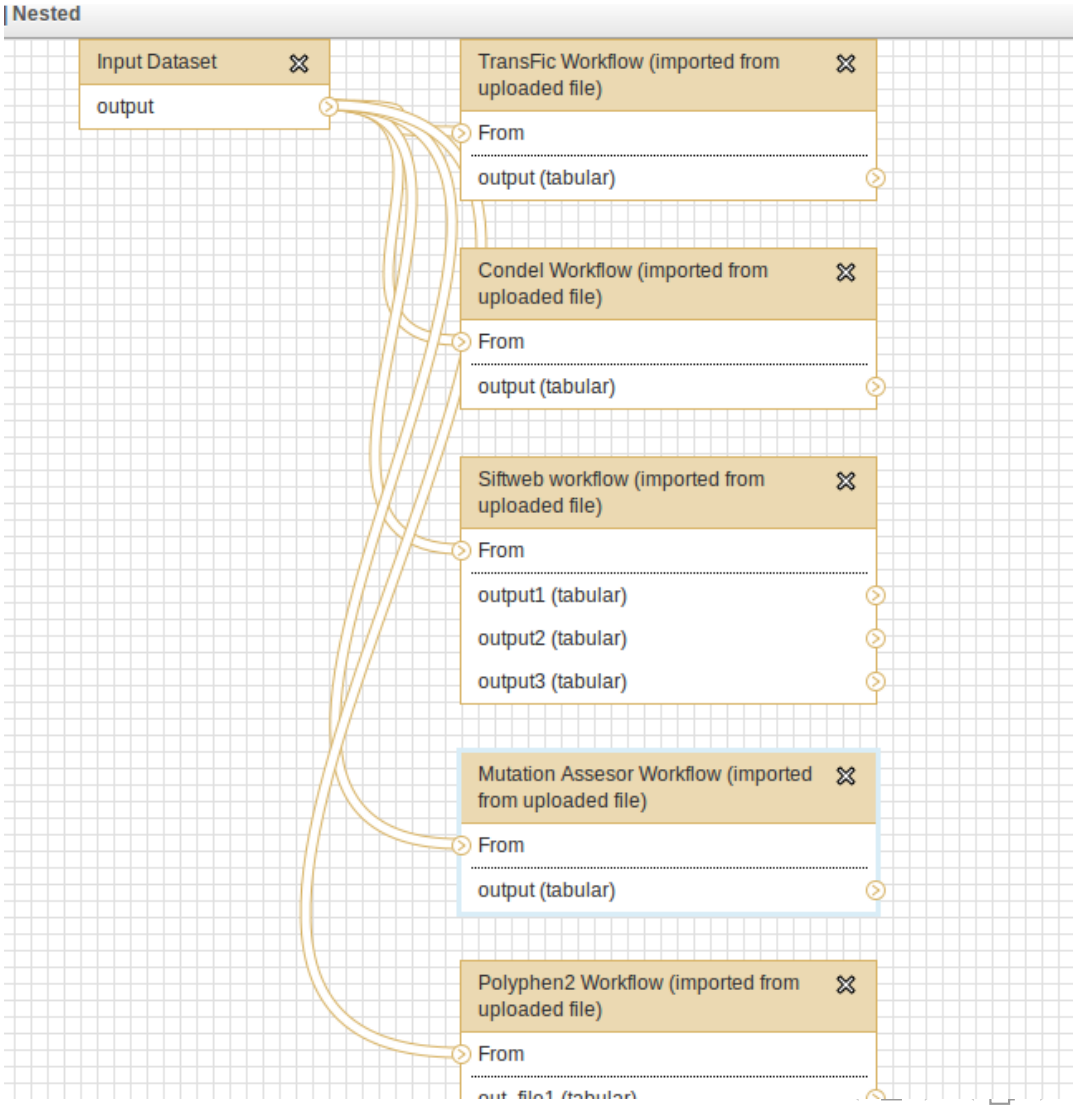
Solution?:
Galaxy[**?**], an open source web-based platform for bioinformatics, makes it possible to represent the entire data analysis pipeline in an intuitive graphical interface

Figure: Galaxy Workflow polyphen2 algorithm



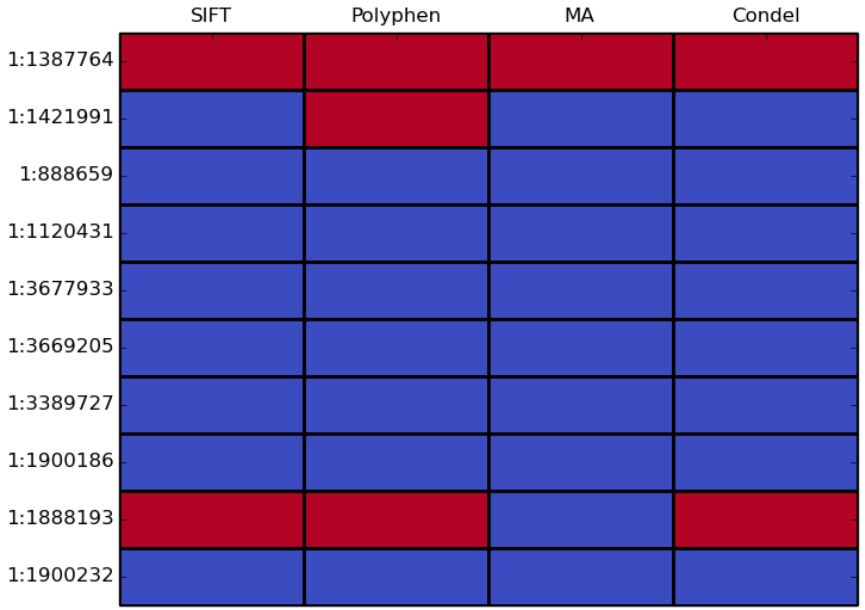
Run all tools in one go:

Figure: Run all tools



Compare all tools:

Figure: Compare all tools



VIRAL GENOME DETECION

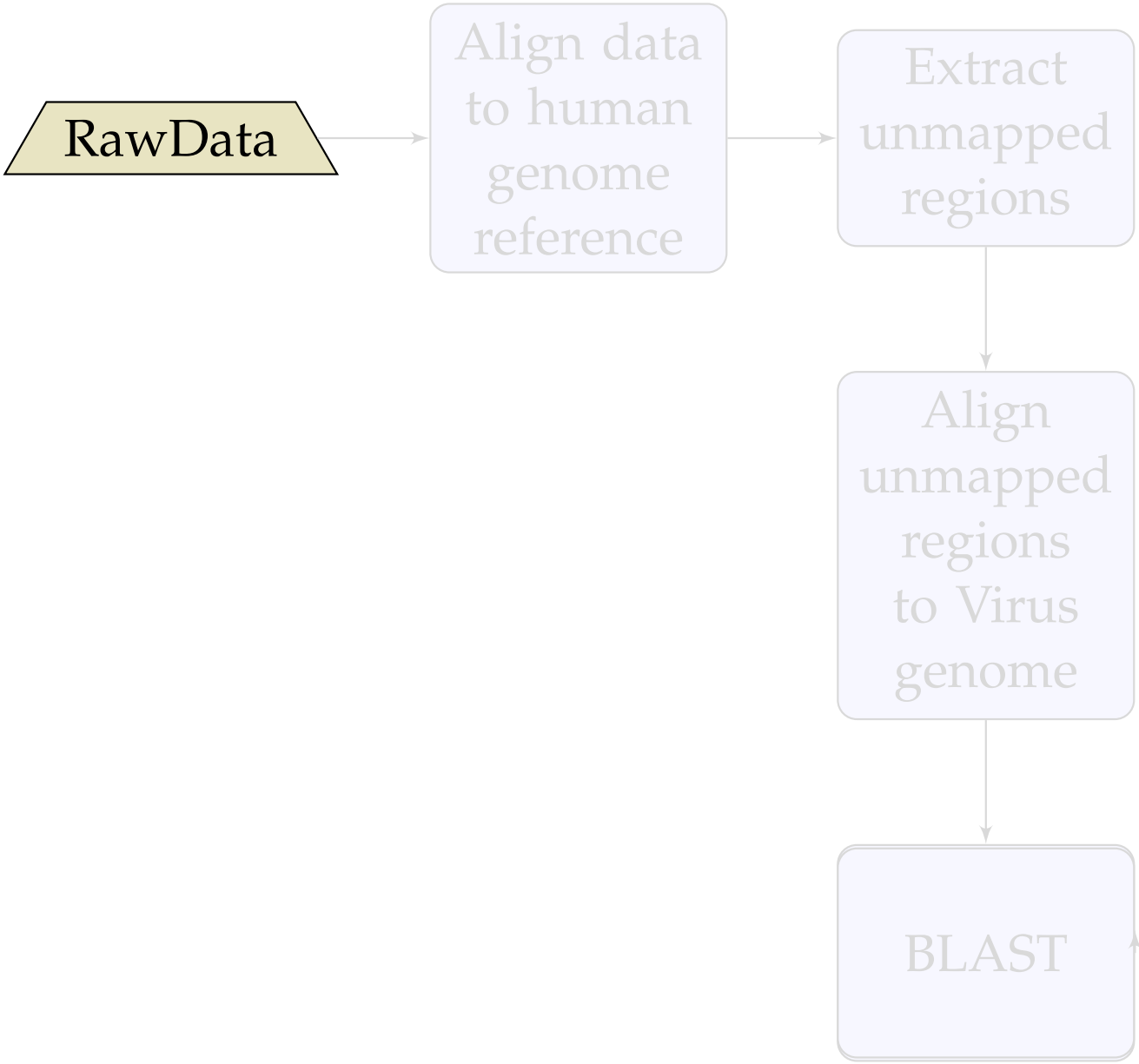
Cervical cancers have been proven to be associated with Human Papillomavirus(HPV)
Cervical cancer datasets from Indian women was put through an analysis to detect :

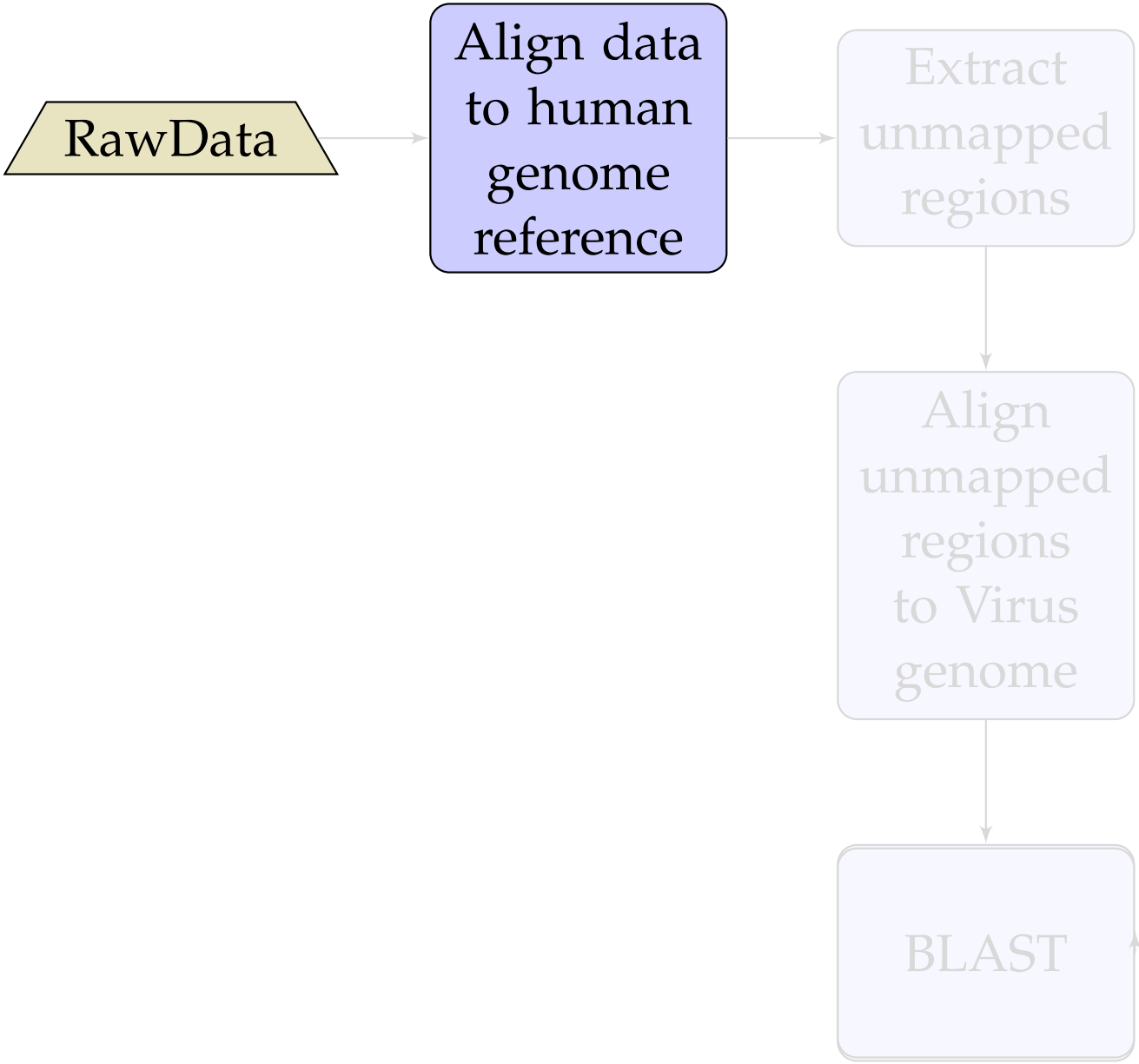
1. Any possible HPV integration
2. Sites of HPV integration

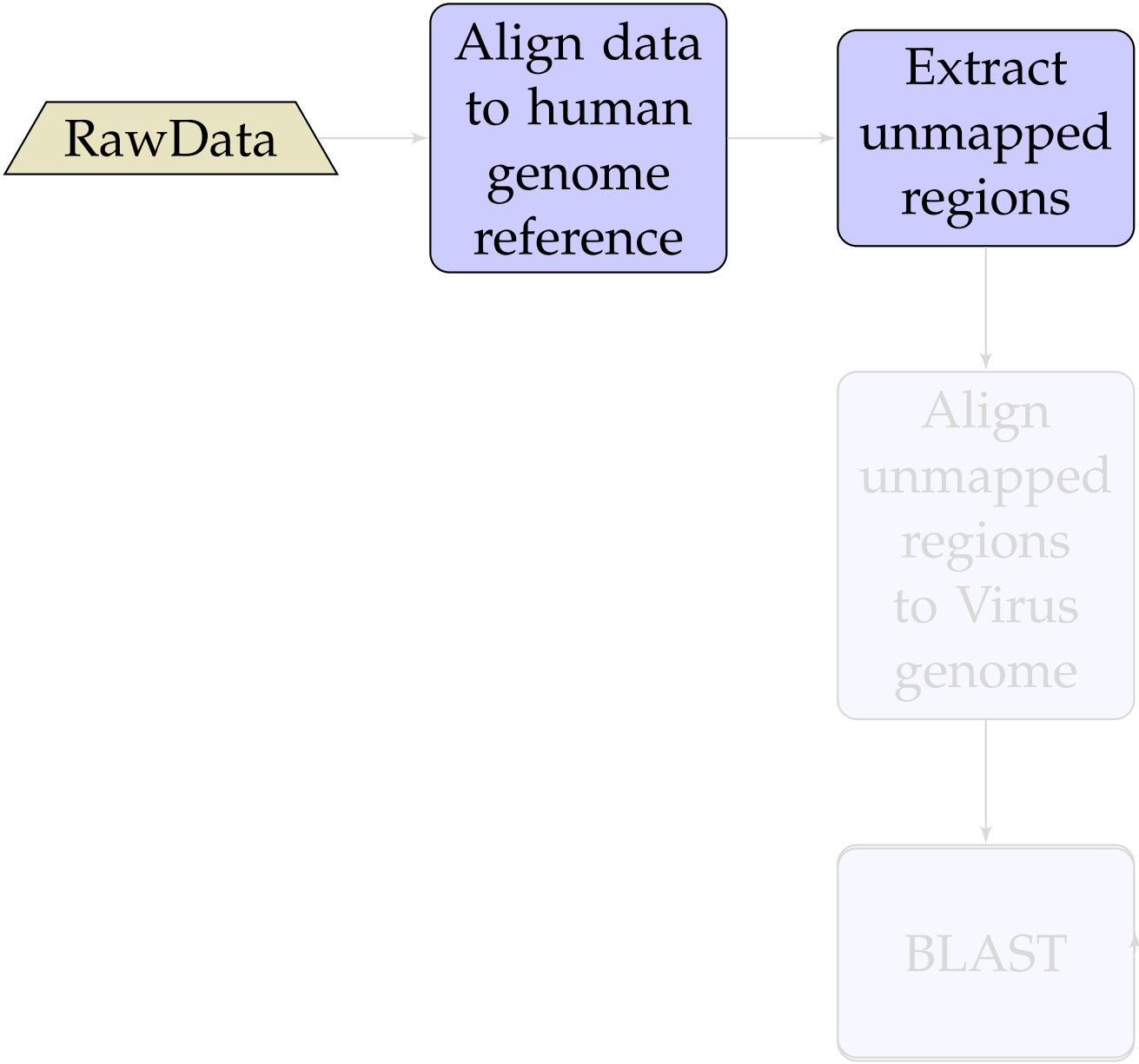
Who Cares?

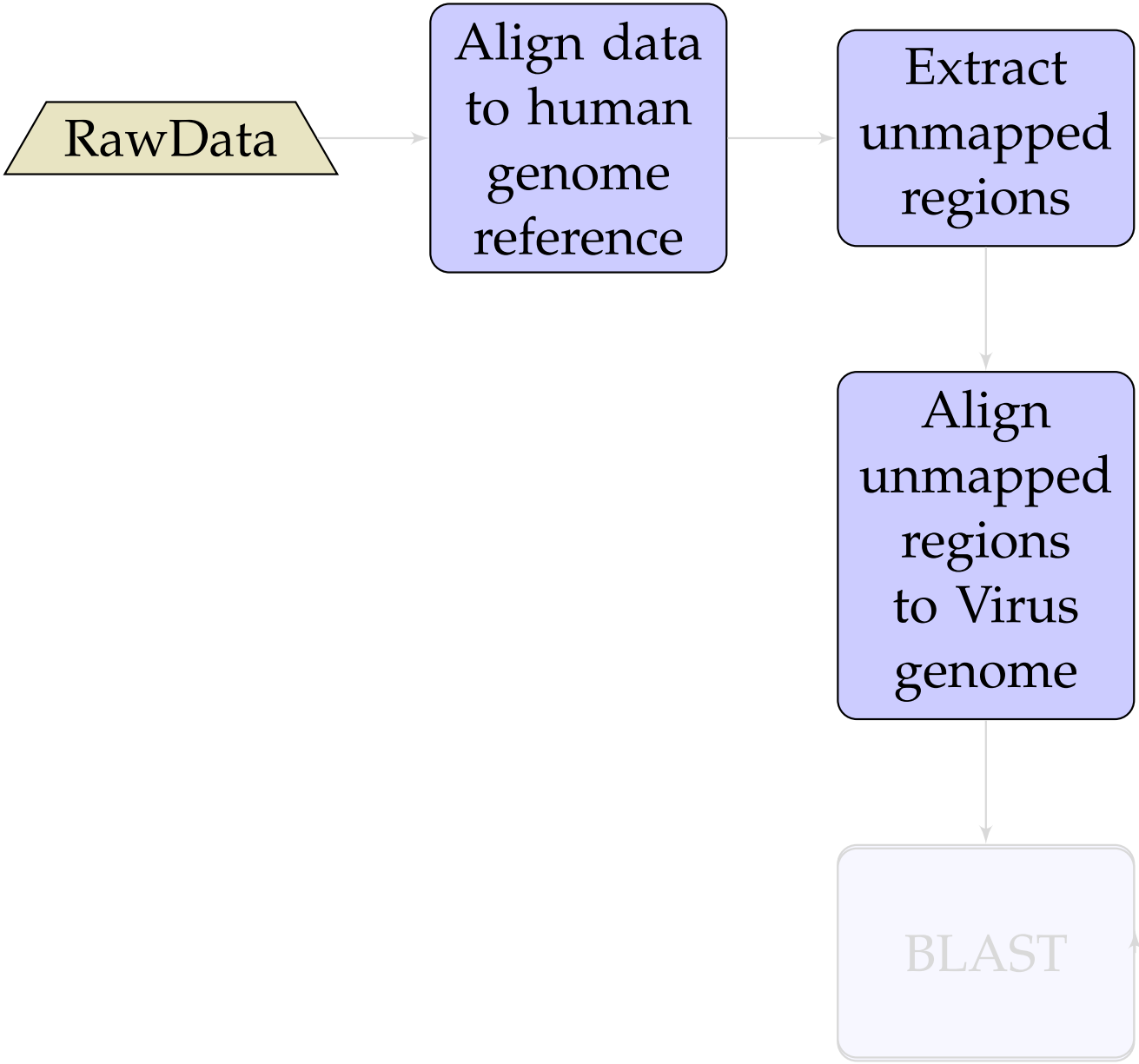
- Replacing whole genome sequencing, by targeted sequencing at the sites where these virus have been detected in a cohort of samples, thus speeding up the whole process.

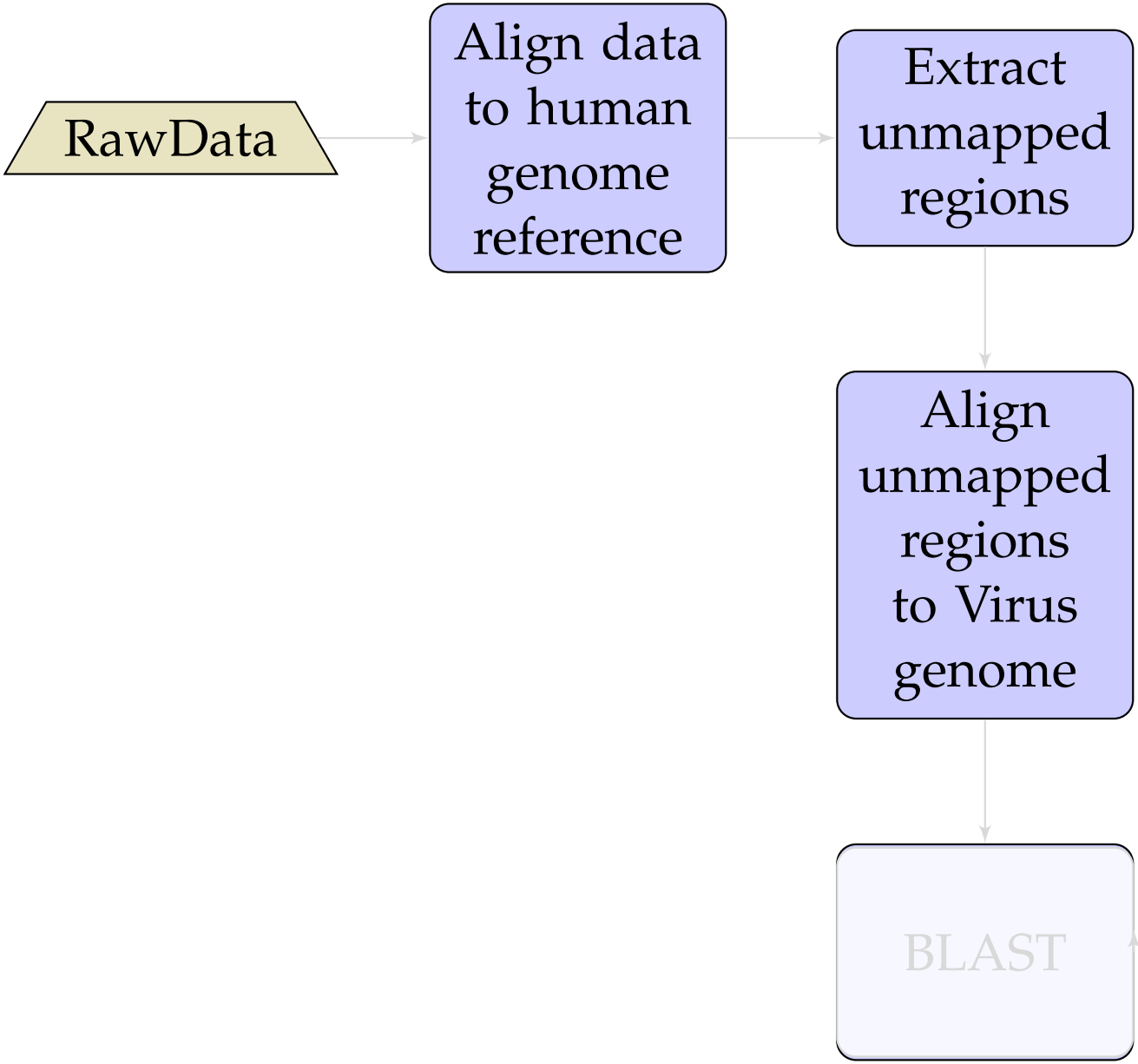
RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST

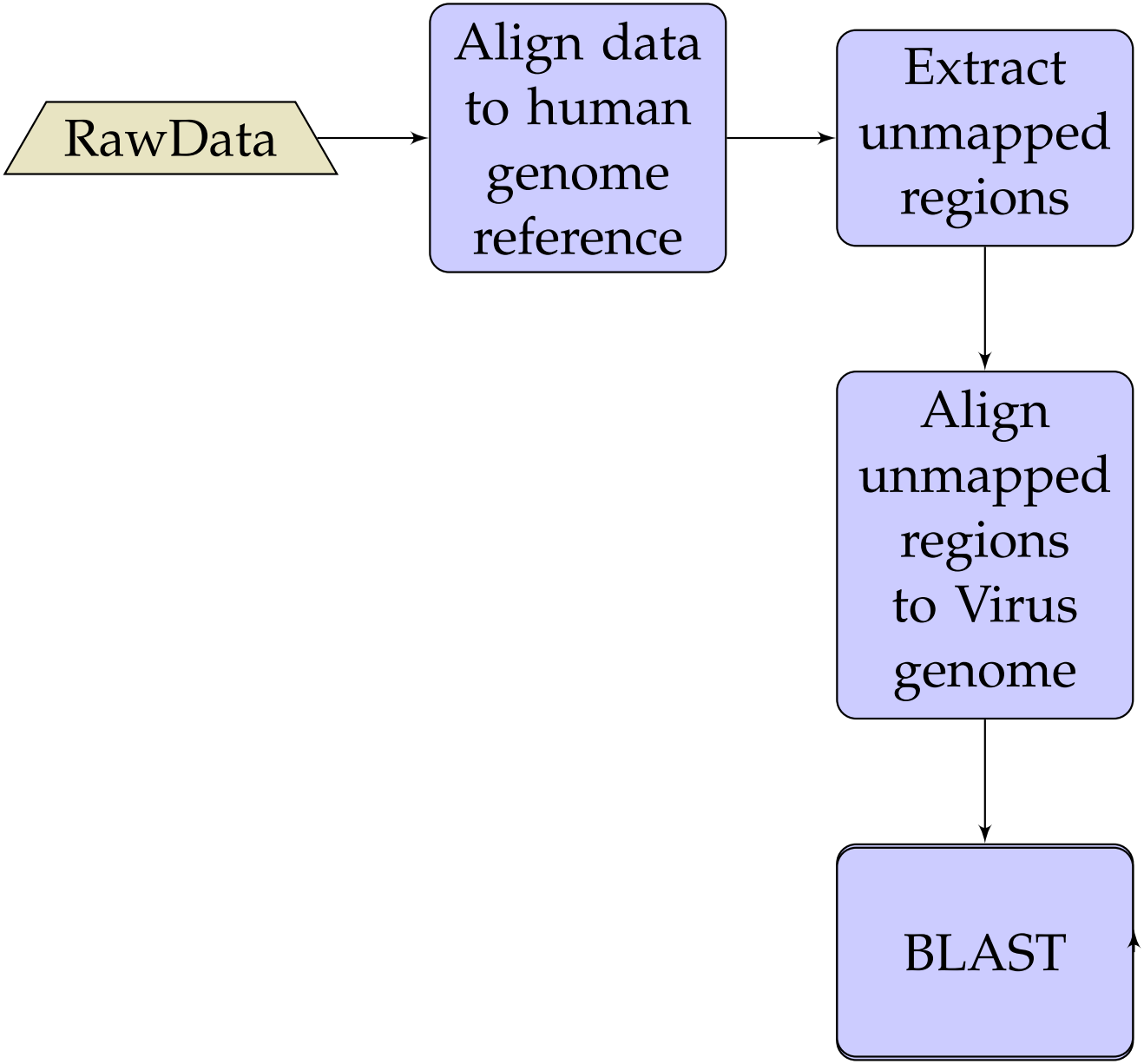

RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST


RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST


RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST


RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST


RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST


RawData
Align data to human genome reference
Extract unmapped regions
Align unmapped regions to Virus genome
BLAST

GALAXY WORKFLOW

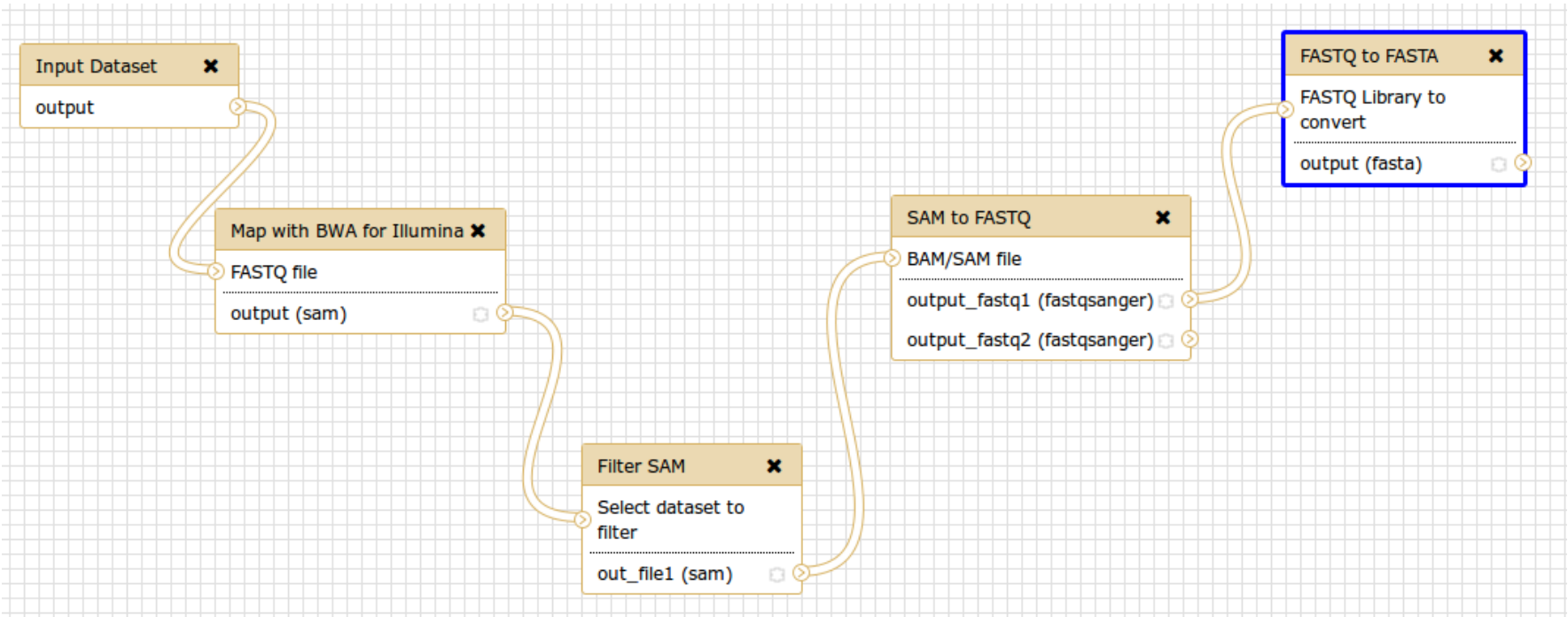
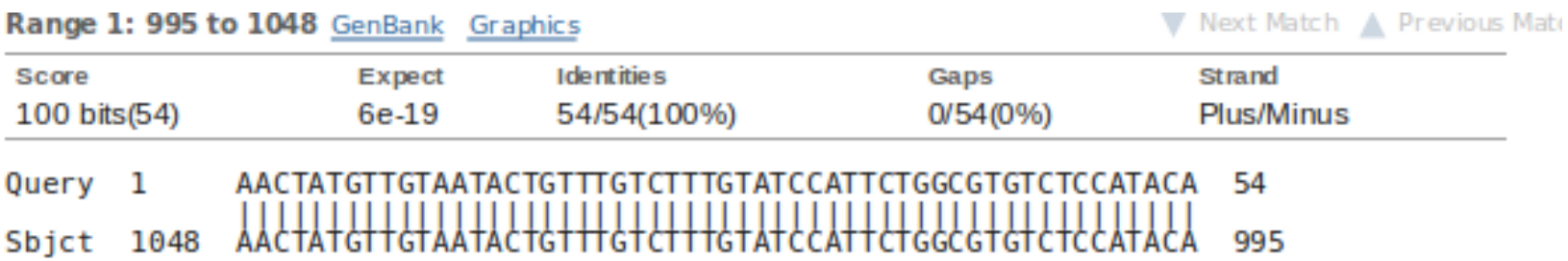


Figure: Aligned HPV genomes



Range 1: 995 to 1048 GenBank Graphics ▼ Next Match ▲ Previous Mat

Score	Expect	Identities	Gaps	Strand
100 bits(54)	6e-19	54/54(100%)	0/54(0%)	Plus/Minus

```
Query  1     AACTATGTTGTAATACTGTTTGTCTTTGTATCCATTCTGGCGTGTCTCCATACA  54
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1048  AACTATGTTGTAATACTGTTTGTCTTTGTATCCATTCTGGCGTGTCTCCATACA  995
```

REPRODUCIBILITY

- In pursuit of novel 'discovery', standardizing the data analysis pipeline is often ignored, leading to dubious conclusions
- Analysis should be reproducible and above all, correct
- Parameter's values can change the results by a big factor, they need to be documented/logged
- *Garbage in, Garbage out*

CONCLUSIONS

With the Galaxy tool box for identification of significant mutations and the study of the science behind the methods, the next steps would be to:

- **Open source the toolbox to the community**: A tool makes little sense if it is not in a usable form, community feedback will be used to add more tools and improve the existing ones
- **A new method for driver mutation prediction**: all the methods have low level of concordance. A new method that takes into account the available data at all levels : mutations, transcriptome and micro array data is possible. With the Galaxy toolbox in place, it would be possible to integrate information at various levels

FUTURE WORK

- Develop an algorithm that integrates machine learning approach with functional approach by zeroing down upon only those attributes that are *known* to have an impact
- The algorithm would also account for information at other levels: RNA expressions, Clinical data.
- Integrating information at all levels would provide a deeper insight
- The developed Galaxy toolbox will be used a the basic framework for integrating information

REFERENCES I

Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin.
Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.
Cancer research, 69(16):6660–6667, 2009.