Bacterial Genomics Workshop

March 28th -30th 2016

Goals of workshop

- Get an overview of steps in microbial genomics pipeline

- Get exposure to common file formats and terminology in genomics

- Get hands on experience with a set of tools that could compose a genomics pipeline

- Get experience working in a high-performance computing environment

Format of workshop sessions

1. Start with an overview of where the current session fits into the larger pipeline and introduce the steps/tools (~10 min)

2. Demonstration of tools and overview of input and output file formats (~10-20 min)

3. Students work through labs to gain hands on experience with data/tools, with instructors on hand to answer questions and troubleshoot problems

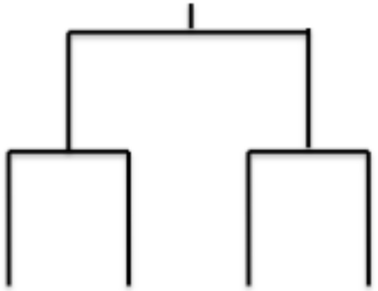
Caveats

- This is the first time we are piloting this material (read – let us know if things are unclear!)

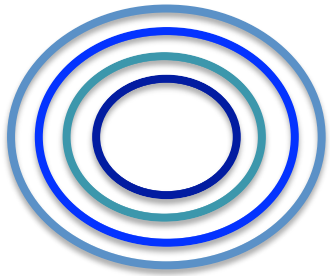
- This is the first time students are going through these lab materials (read – there may be some bugs ☺)

So you want to sequence some bacteria?

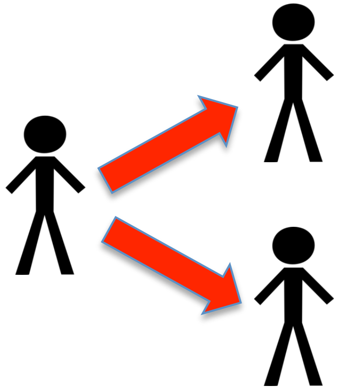
- Microbial phylogenetics



- Comparative genomics

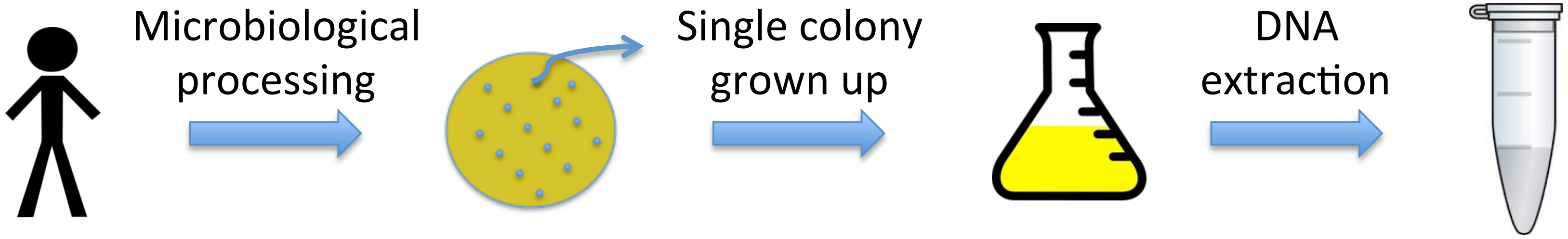


- Genomic epidemiology



DNA and library preparation

1. Sample Preparation

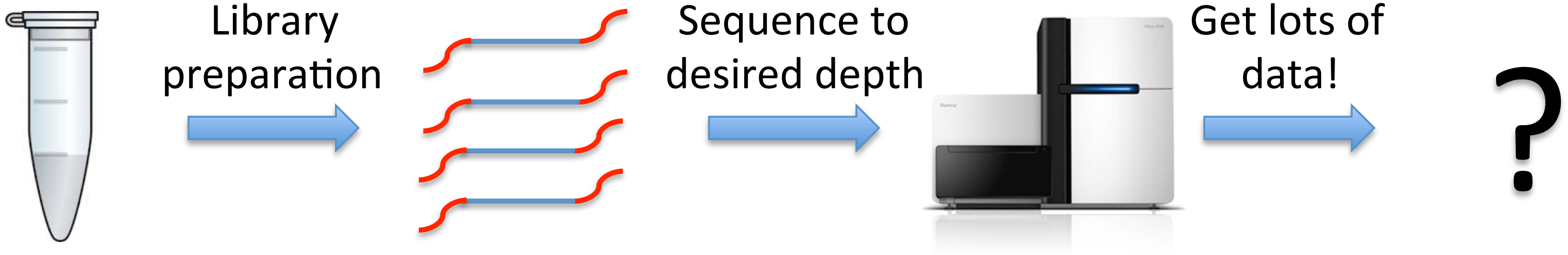


Microbiological processing

Single colony grown up

DNA extraction

2. Sequencing



Library preparation

Sequence to desired depth

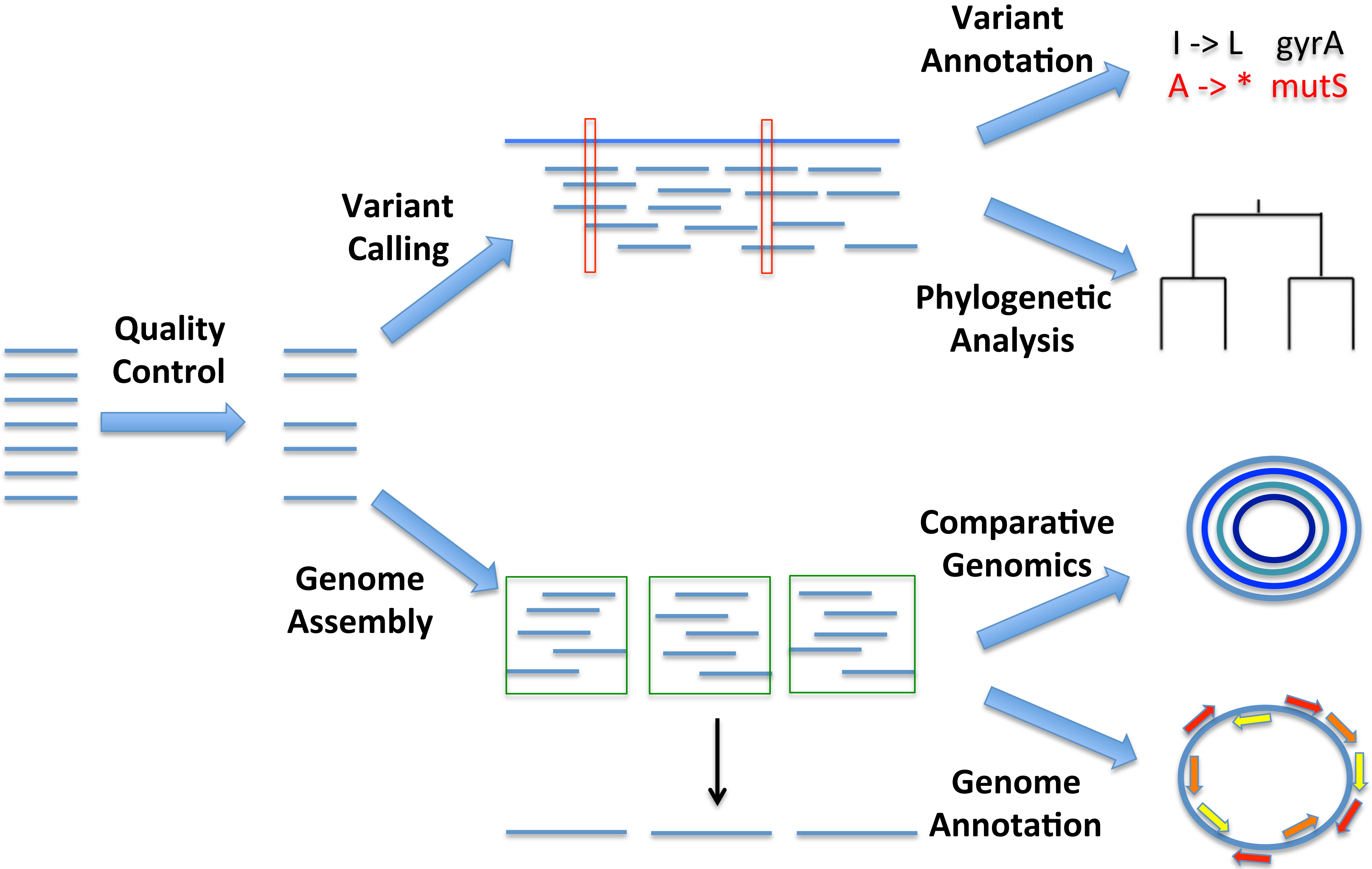
Get lots of data!

?

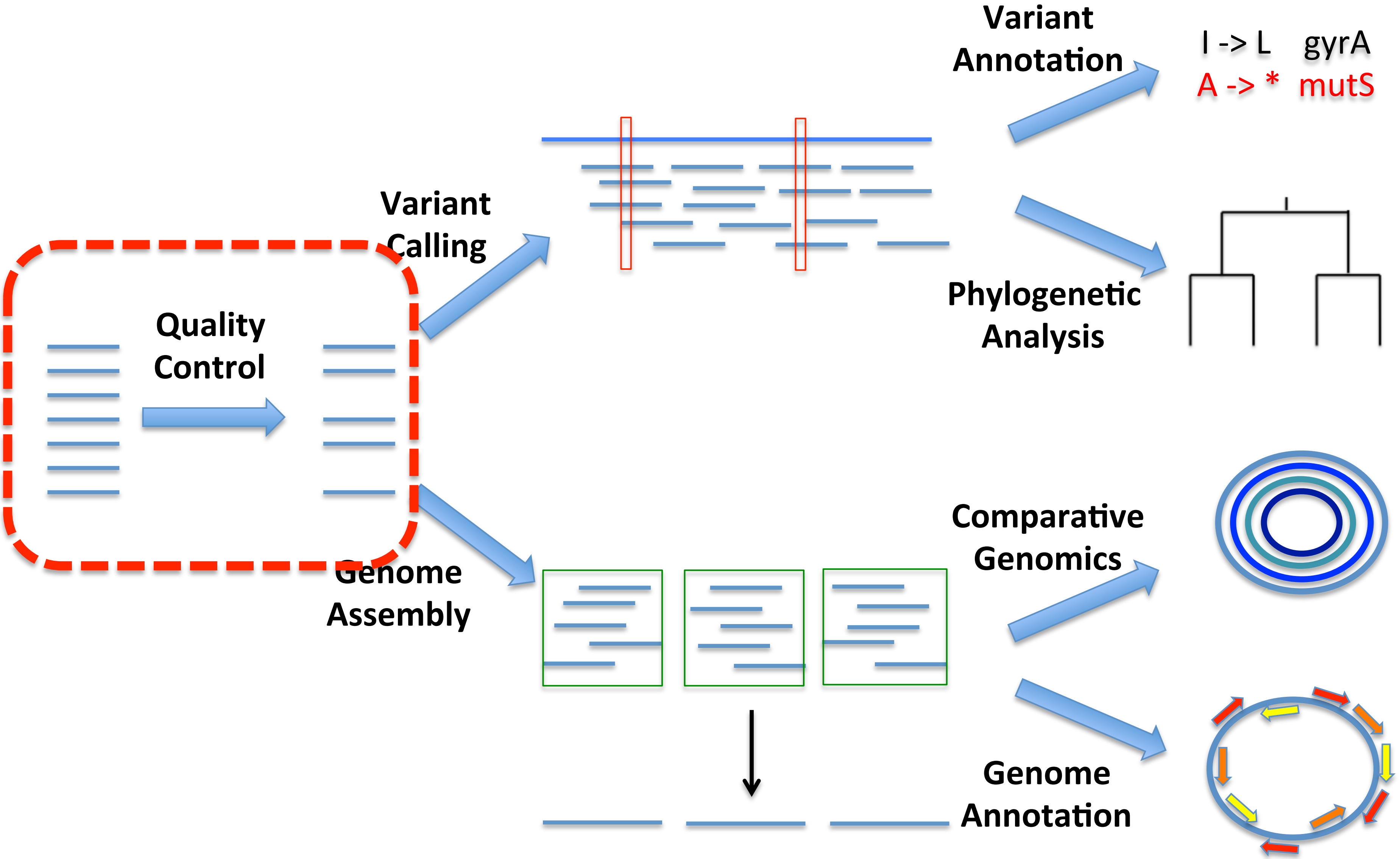
Illumina sequencing

- https://youtu.be/womKfikWlxM

Mile-high view of a genomics pipeline



Mile-high view of a genomics pipeline



Sequencing quality control

Forward reads

```
@seq1_1
ACTGCACT
+
8-8,,+@+
@seq2_1
TGCATCTA
+
@+@E++BF
.
.
.
```

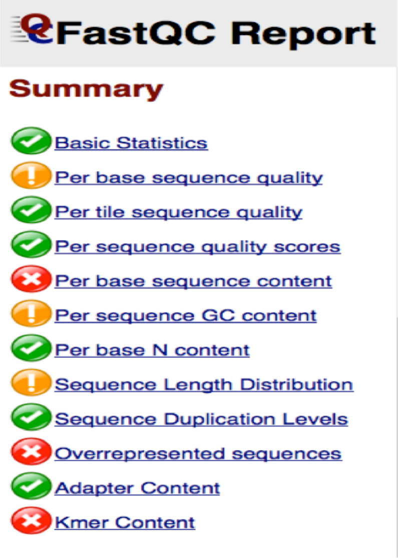
Reverse reads

```
@seq1_2
TCTATCGA
+
A<-9BFCFF
@seq2_2
CTAGTTAA
+
**>D7?7=.
.
.
.
```

Raw fastq files

FastQC

1. Contaminants
2. Aberrant quality



FastQC Report

Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

Trimmomatic

1. Filter reads
2. Trim reads

Forward reads

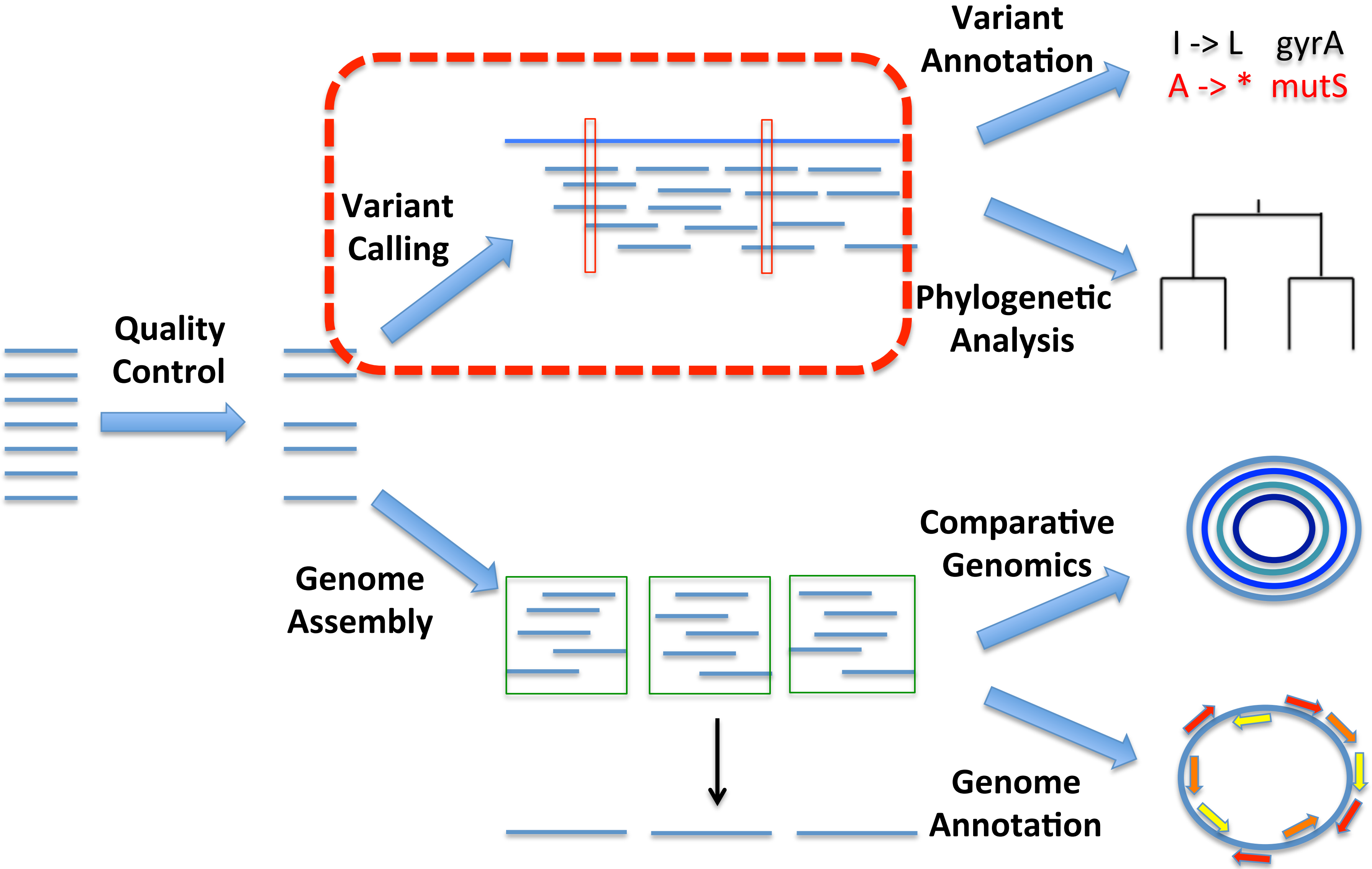
```
@seq1_1
ACTGCACT
+
8-8,,+@+
.
.
.
.
.
.
```

Reverse reads

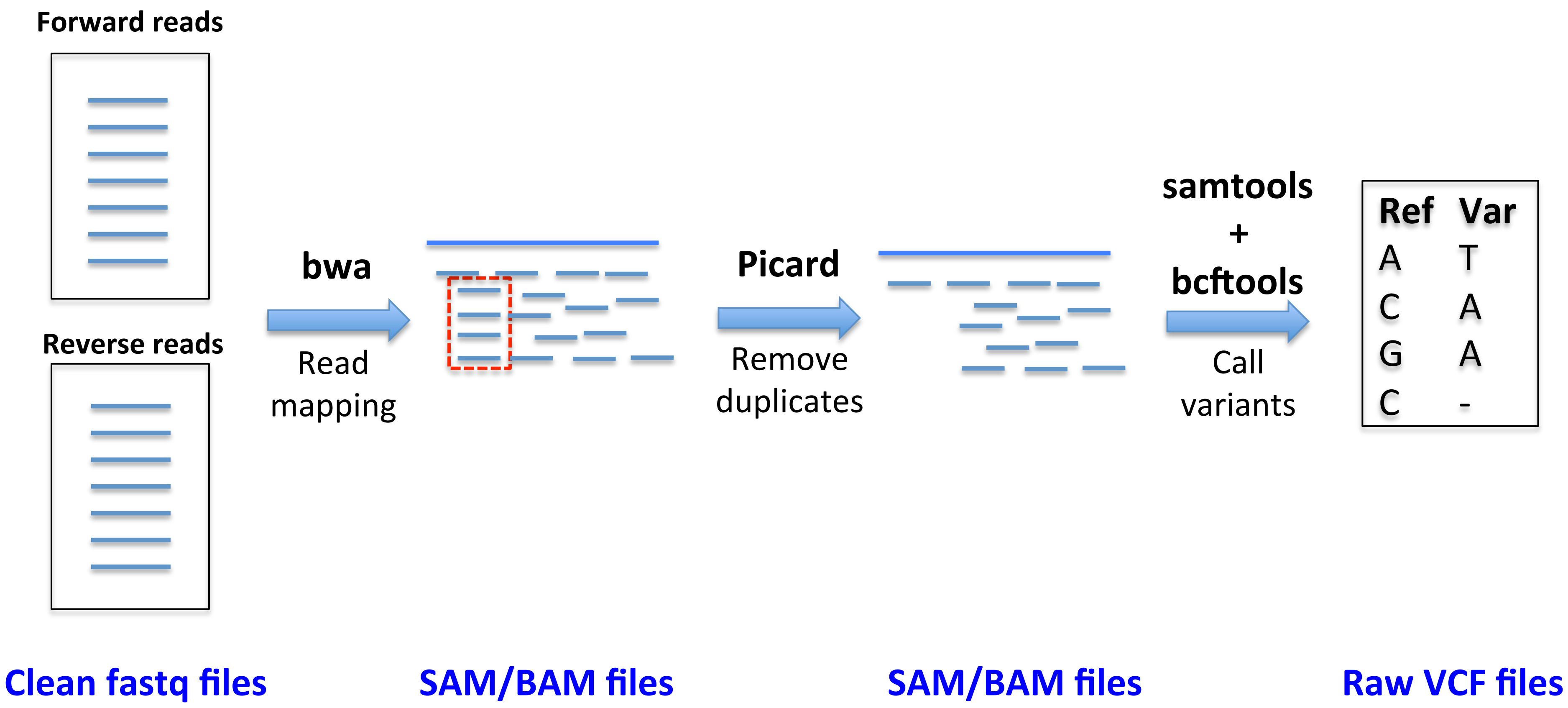
```
@seq1_2
TCTATCGA
+
A<-9BFCFF
.
.
.
.
.
.
```

Clean fastq files

Mile-high view of a genomics pipeline



Variant identification



Forward reads

Reverse reads

bwa

Read mapping

Picard

Remove duplicates

samtools + bcftools

Call variants

Ref	Var
A	T
C	A
G	A
C	-

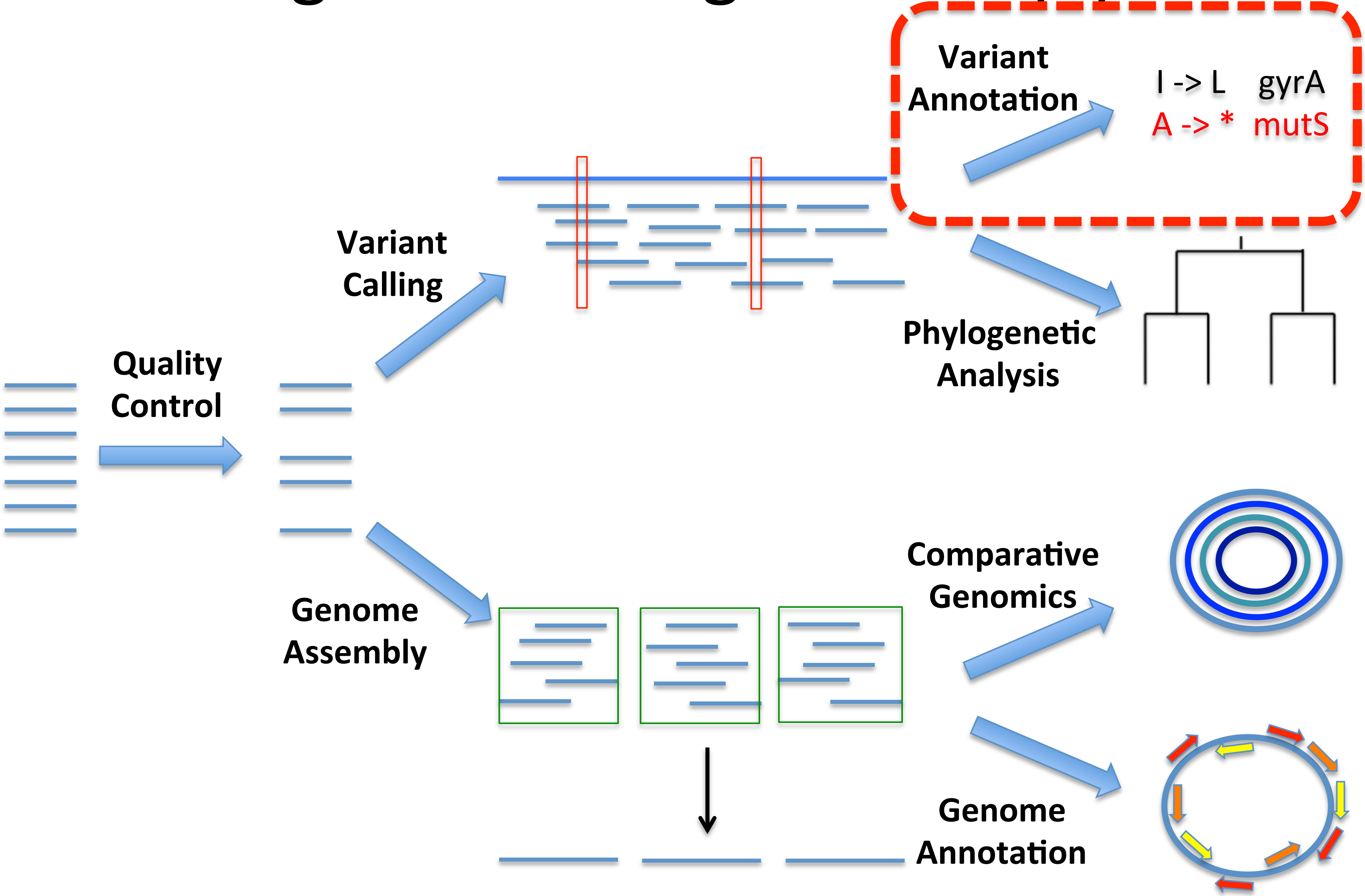
Clean fastq files

SAM/BAM files

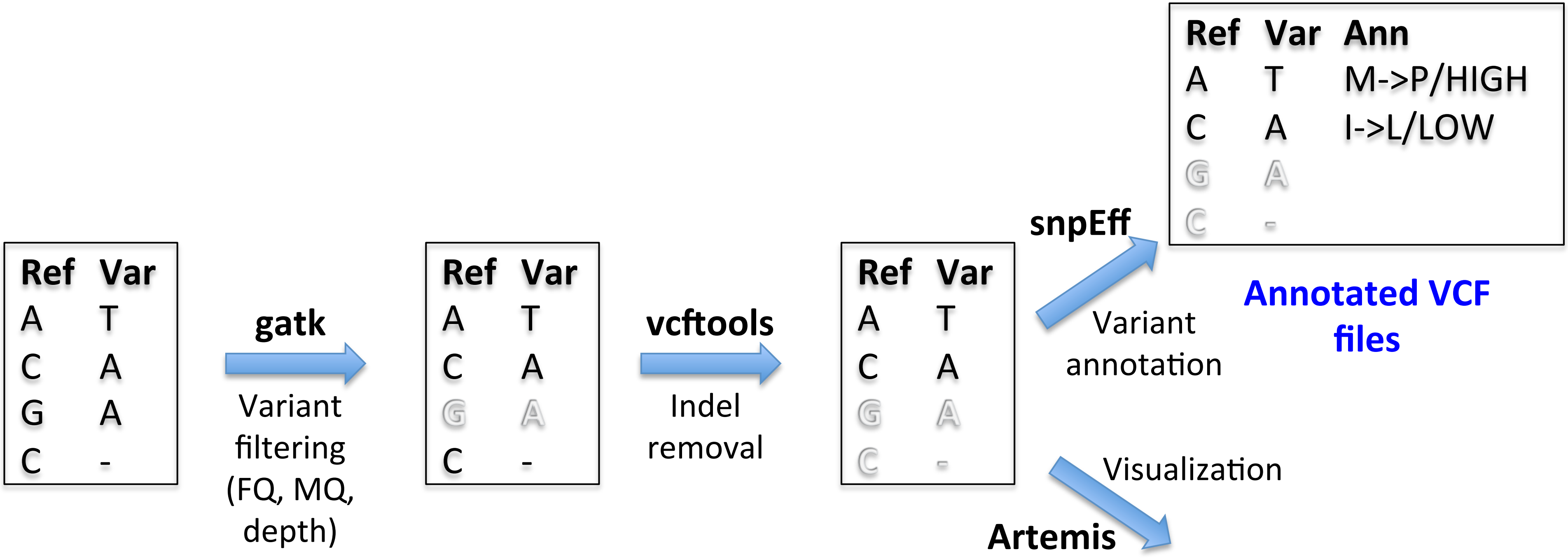
SAM/BAM files

Raw VCF files

Mile-high view of a genomics pipeline



Variant filtering and annotation



Ref	Var
A	T
C	A
G	A
C	-

gatk

Variant filtering (FQ, MQ, depth)

Ref	Var
A	T
C	A
G	A
C	-

vcftools

Indel removal

Ref	Var
A	T
C	A
G	A
C	-

snpEff

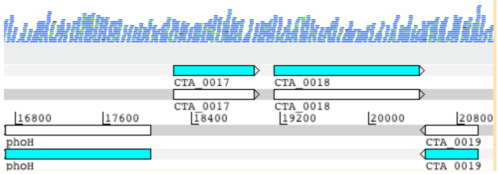
Variant annotation

Ref	Var	Ann
A	T	M->P/HIGH
C	A	I->L/LOW
G	A	
C	-	

Annotated VCF files

Visualization

Artemis



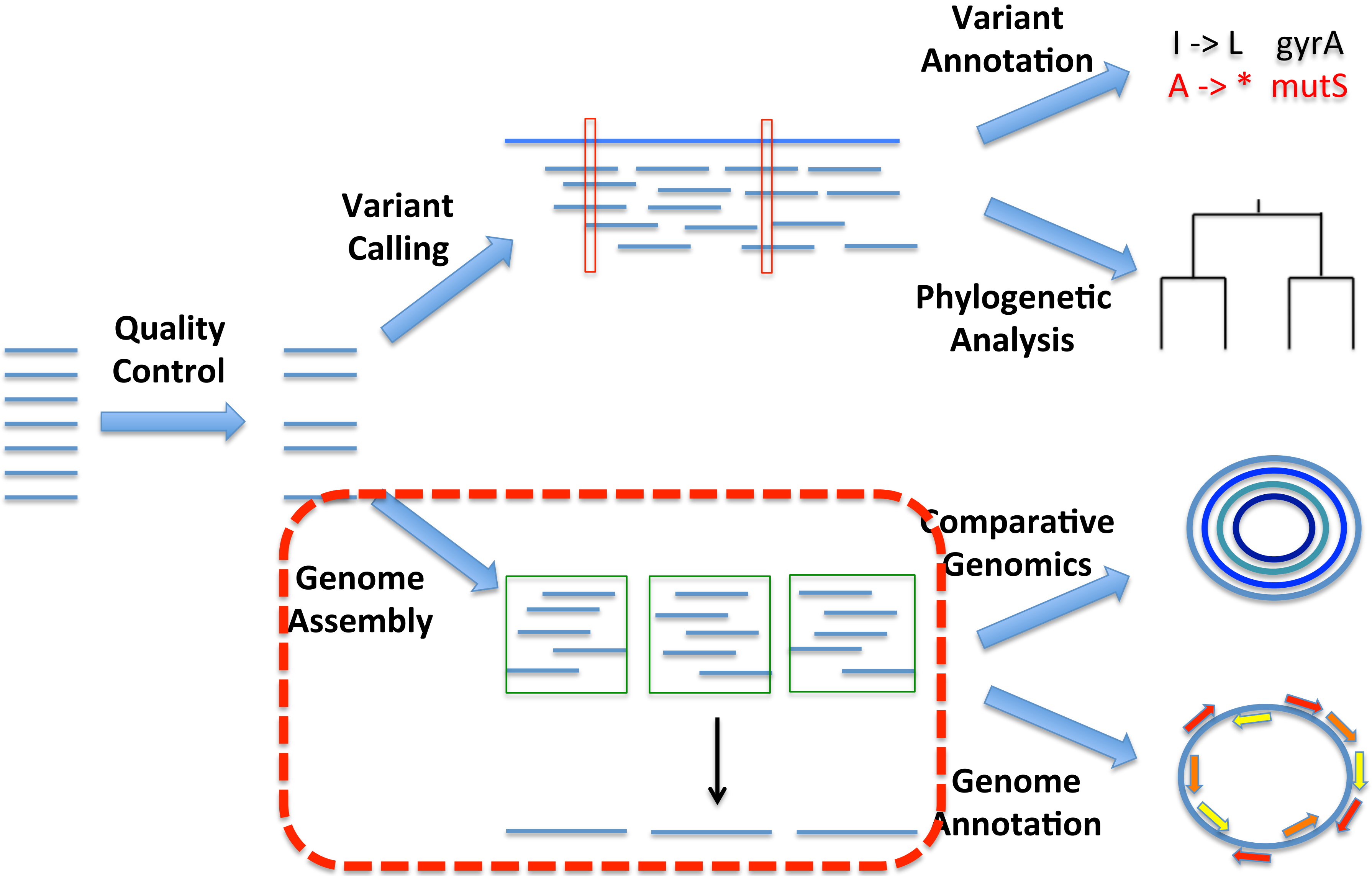
VCF files

Filtered VCF files

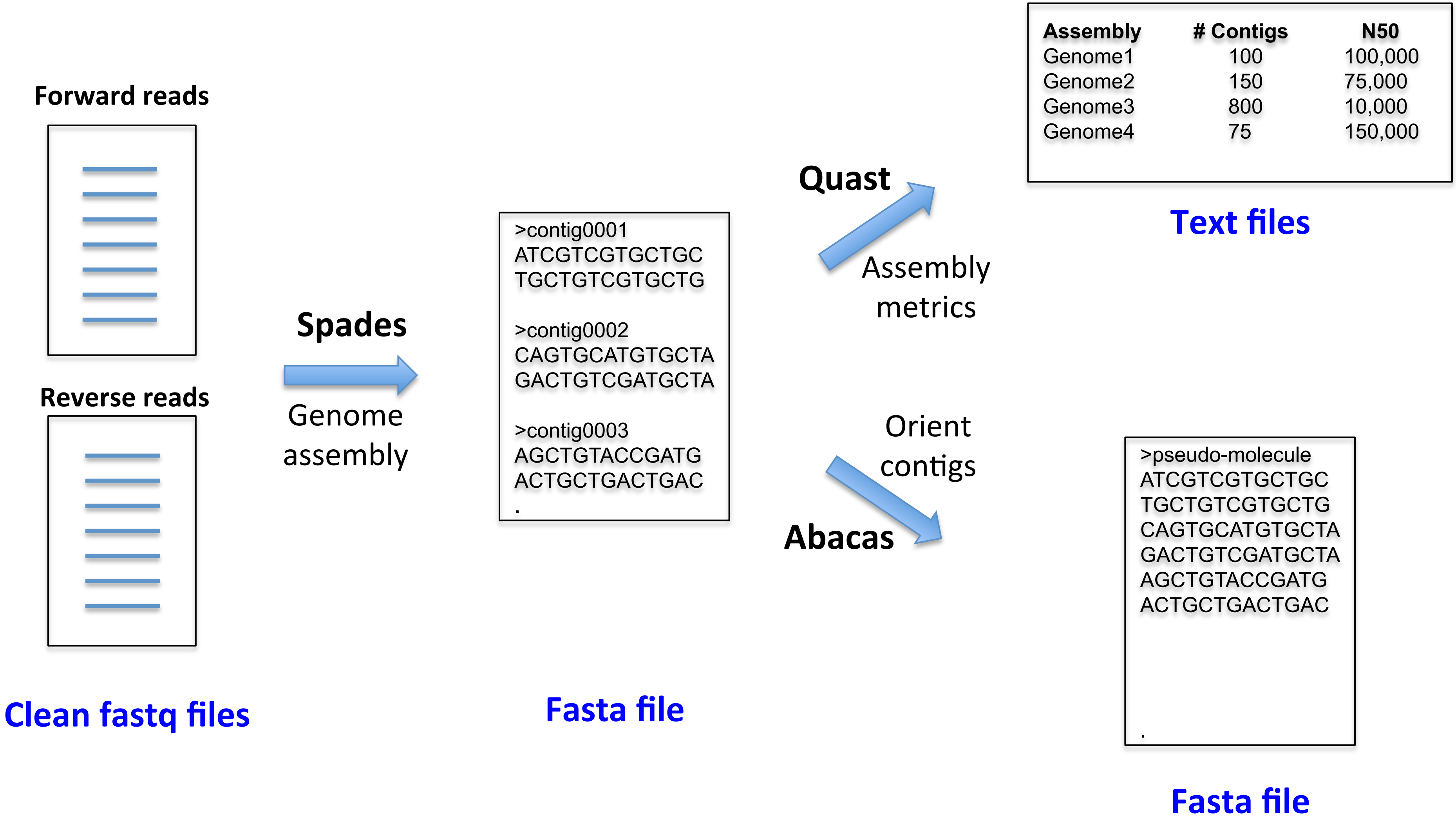
Filtered VCF files

VCF, BAM, BAI, fasta files

Mile-high view of a genomics pipeline



Genome assembly



Forward reads

Reverse reads

Clean fastq files

Spades

Genome assembly

```
>contig0001
ATCGTCGTGCTGC
TGCTGTCGTGCTG

>contig0002
CAGTGCATGTGCTA
GACTGTCGATGCTA

>contig0003
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Fasta file

Quast

Assembly metrics

Assembly	**# Contigs**	**N50**
Genome1	100	100,000
Genome2	150	75,000
Genome3	800	10,000
Genome4	75	150,000

Text files

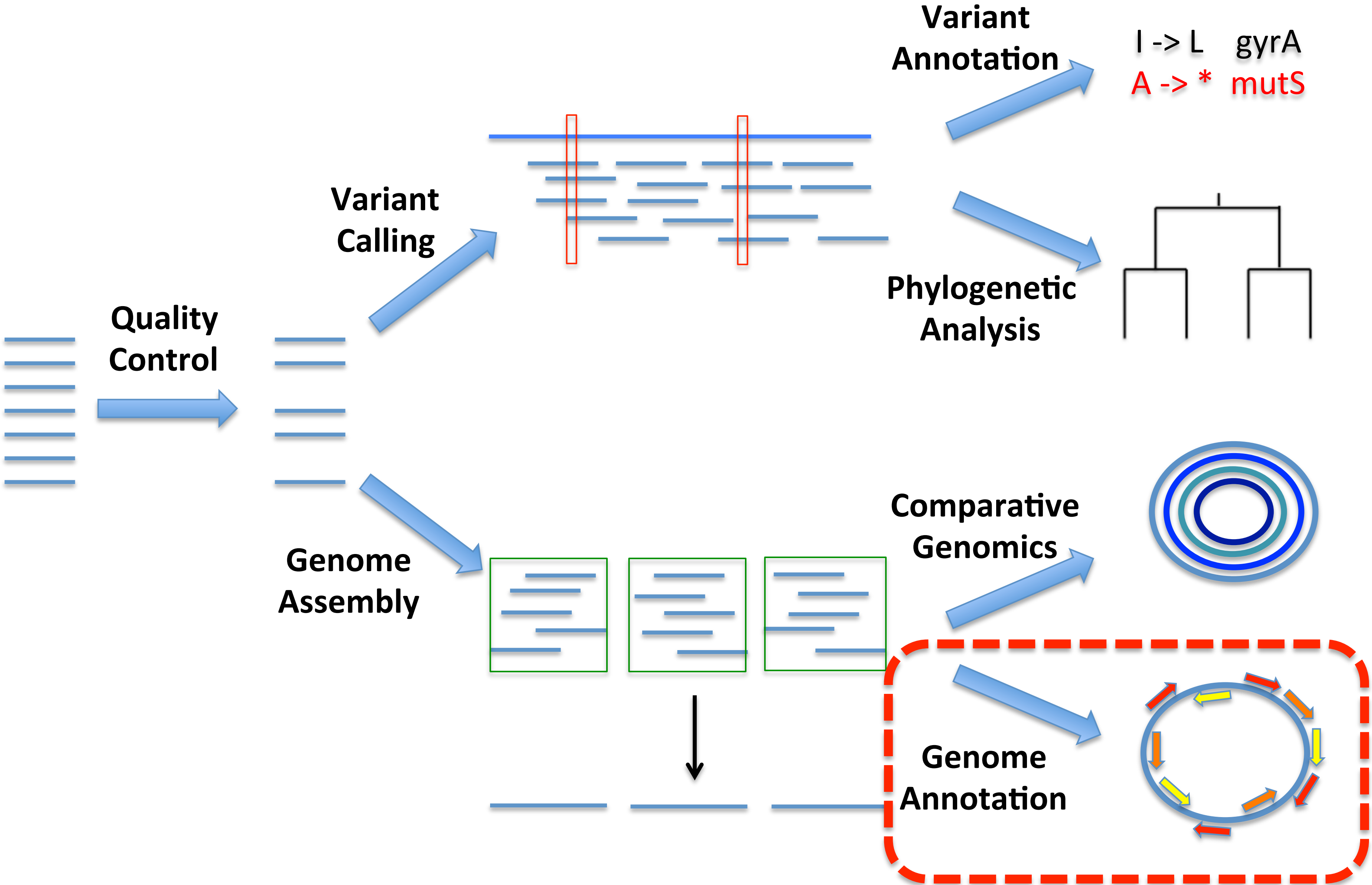
Orient contigs

Abacas

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCTA
GACTGTCGATGCTA
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Fasta file

Mile-high view of a genomics pipeline

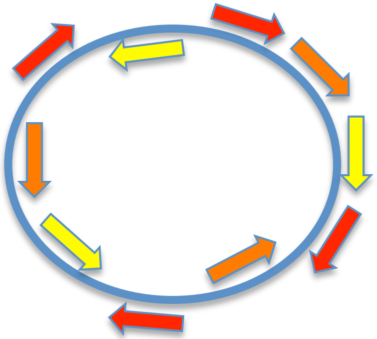


Genome annotation

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCTA
GACTGTCGATGCTA
AGCTGTACCGATG
ACTGCTGACTGAC

.
```

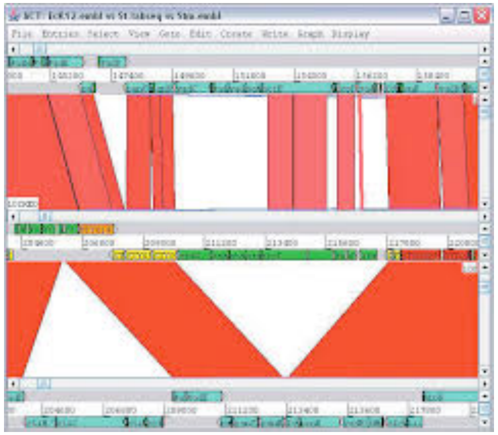
Prokka

1) Gene finding
2) Basic annotation



ACT

Visualization

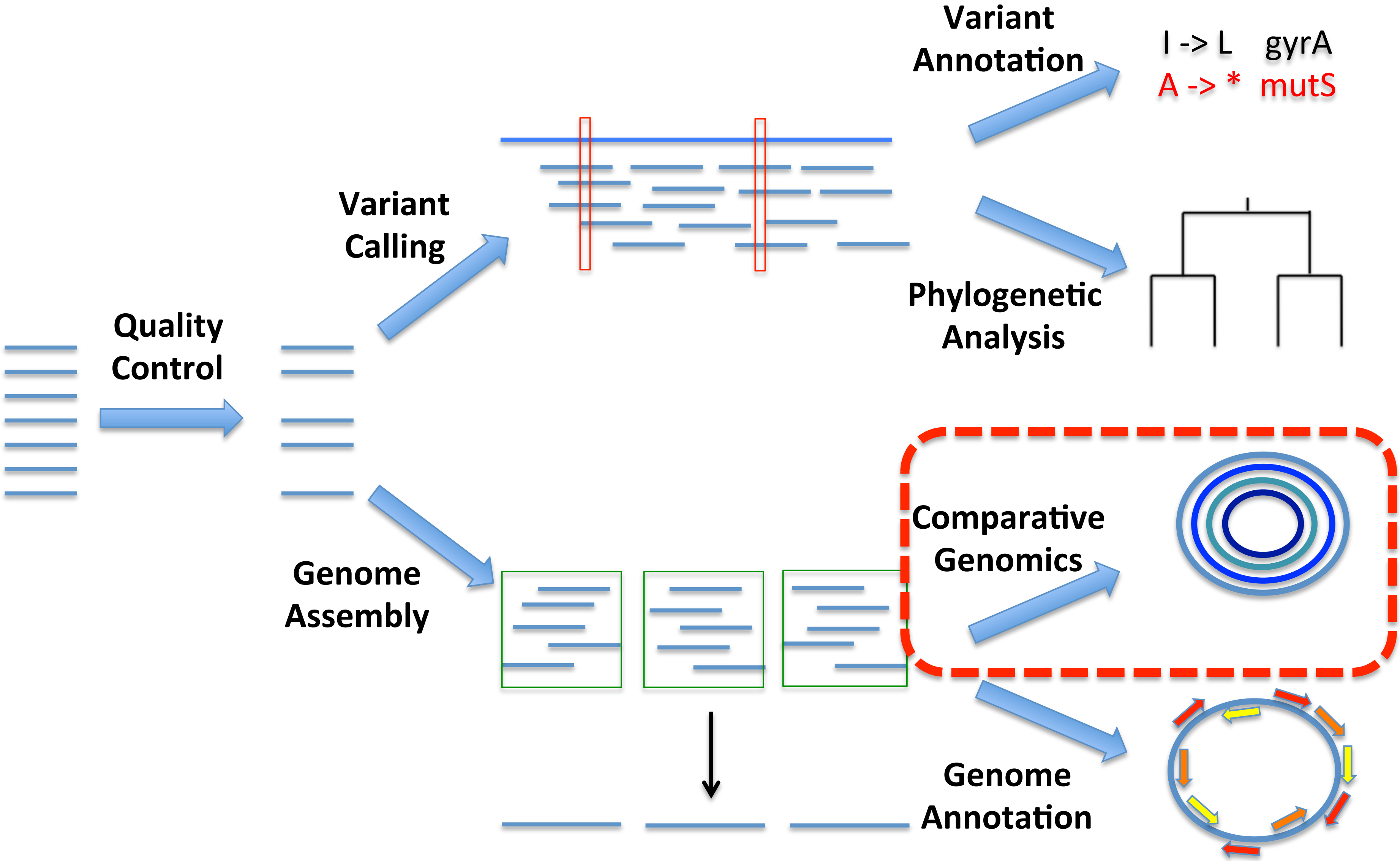


Fasta file

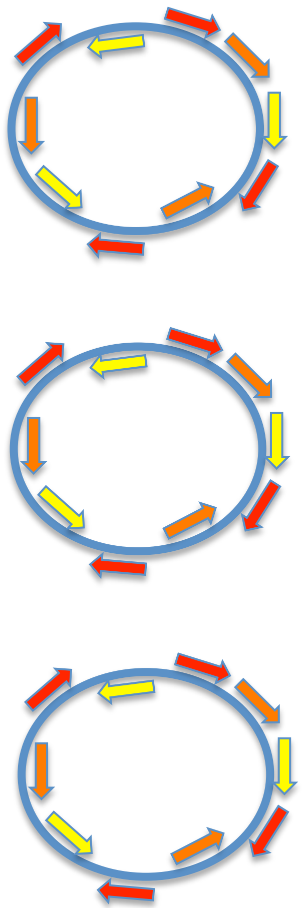
Genbank file

Genbank files, alignment files

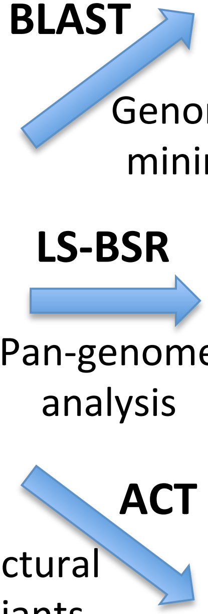
Mile-high view of a genomics pipeline



Comparative genomics



Fasta, genbank and/or pep



BLAST

Genome mining

LS-BSR

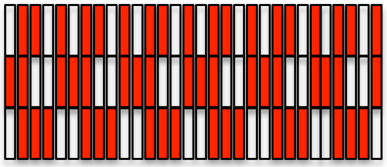
Pan-genome analysis

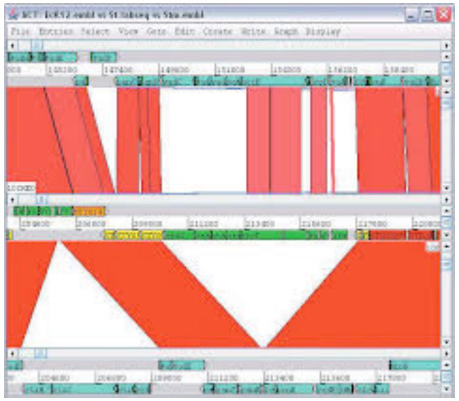
ACT

Structural variants

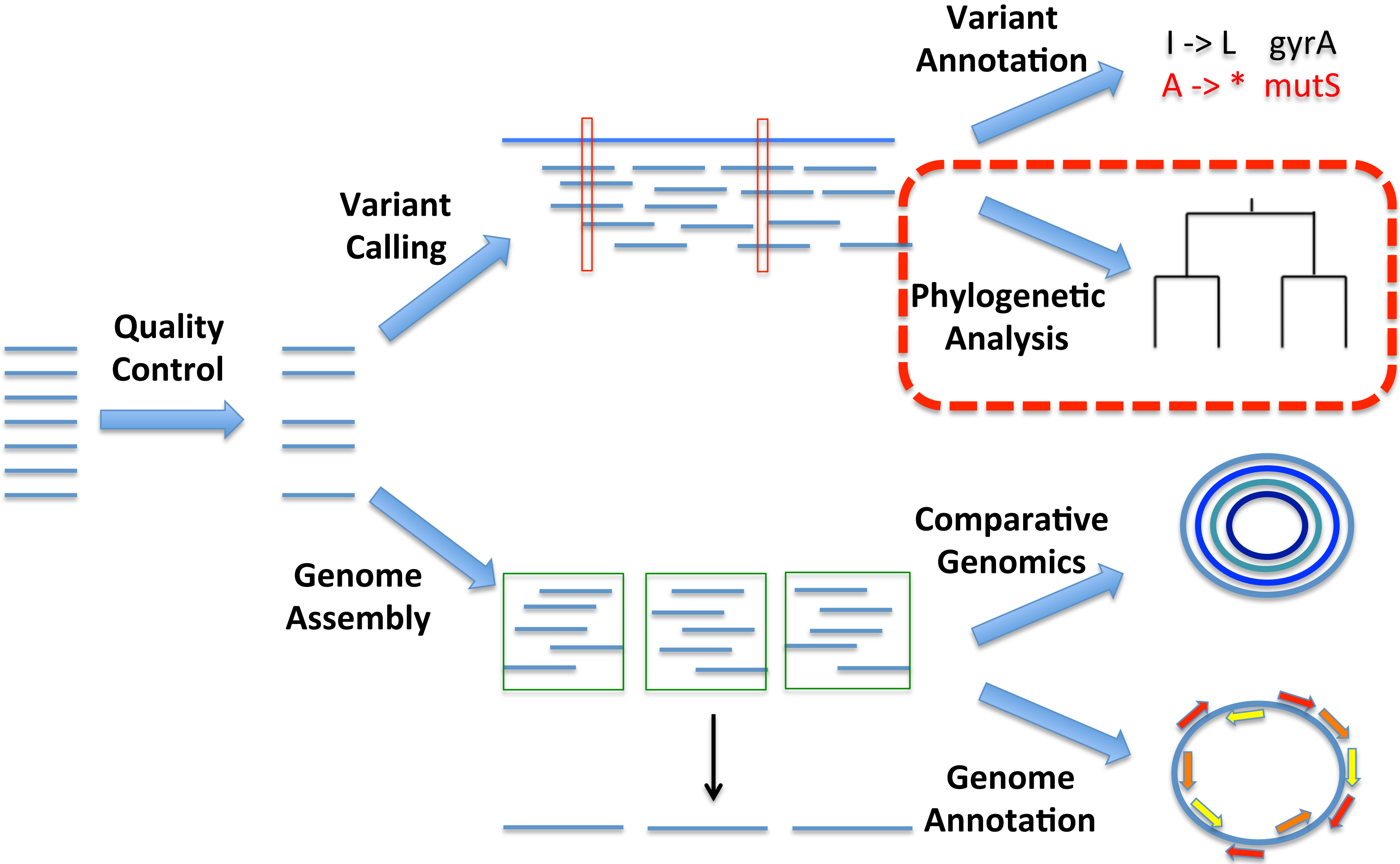
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Genome 1	□	🟥	🟥	□	🟥
Genome 2	🟥	🟥	□	□	🟥
Genome 3	□	🟥	🟥	🟥	□

Genome 1
Genome 2
Genome 3





Mile-high view of a genomics pipeline



Quality Control

Variant Calling

Genome Assembly

Variant Annotation

I -> L gyrA

A -> * mutS

Phylogenetic Analysis

Comparative Genomics

Genome Annotation

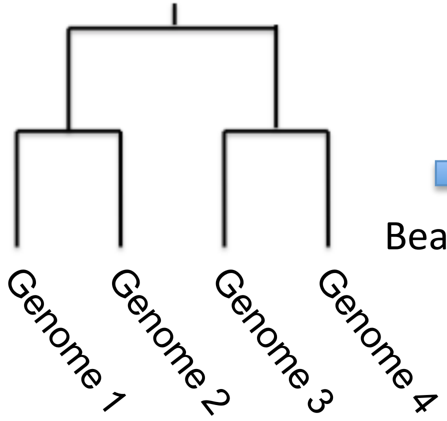
Phylogenetics

```
>Genome 1
ATCGTCGTGCTGC
TGCTGTCGTGCTG

>Genome 2
CAGTGCATGTGCTA
GACTGTCGATGCTA

>Genome 3
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Multi-fasta file

Seaview / ape

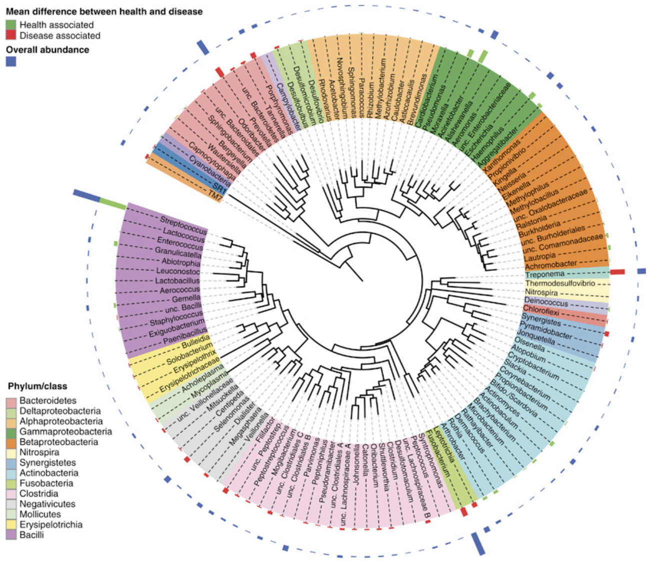
Tree construction



Nexus file

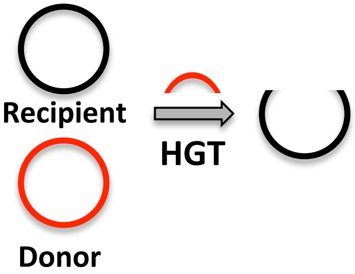
iTOL

Beautification



Gubbins

Recombination filtering



Seaview /ape

Tree construction